QUALITY CONTROL FOR TRANSLATIONAL BIOMEDICAL INFORMATICS

A Dissertation
Presented to
The Academic Faculty

by

Richard Austin Moffitt

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioengineering in the
School of Biomedical Engineering

Georgia Institute of Technology

August, 2009

QUALITY CONTROL FOR TRANSLATIONAL BIOMEDICAL INFORMATICS

Approved by:

Dr. May D Wang, Advisor
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Brian Leyland-Jones
Director of Winship Cancer Institute
Emory University

Dr. Robert Butera
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Andrew N Young
Pathology & Laboratory Medicine
Emory University

Dr. Shuming Nie
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Date Approved: [June 25, 2009]

ACKNOWLEDGEMENTS

Much of the work discussed in this dissertation has been aided by the work of others, and I would like to take this space to thank them all for their help.

Dr. Todd Stokes was responsible for making caCORRECT into the functional web tool and caBIG grid service that it is. Without Todd, caCORRECT would still be a bunch of MATLAB scripts on my desktop PC. Todd is also responsible for coming up with concept of individual chip quality scoring which was prominently featured in his thesis work.

Dr. John Phan was responsible for coding up the currently used C implementation of my artifact-aware quantile normalization scheme. John is also responsible for the current parallel implementation of Support Vector Machines which I have often used to quickly rank features from a microarray platform. Working with John on SVM gene ranking early in my career, and noticing poor reproducibility, is what pushed me to investigate microarray quality control in the first place.

Qiqin Yin-Goen and Dr. Young's lab provided indispensable help for all of the PCR experiments. Qiqin selected all the primers, preformed all RNA extractions from Dr Young's tissue samples, and basically taught me how to do PCR. Dr Young's group provided all domain knowledge on RCC as well as all tissue specimens and microarray data for the RCC study.

Dr. Jian Liu. Dr. Tao Liu and Dr. Nie's lab have provided all quantum dot immunohistochemistry support. Jian was responsible for selecting antibodies to my proteins of interest, conjugating them to quantum dots, and staining all of the tissue slides. Tao performed most of the staining of slides for the QD-calibration studies.

Matt Caldwell is an undergraduate mentee that has been helping me with collecting, managing, and processing all of our QD imaging data. I'd like to thank Matt in particular for being the main force behind the digitization of most of the QD data shown here.

JT Torrance is a formerly undergraduate, now graduated, mentee of mine who has been a fantastic help throughout much of my work. JT started out by searching and

curating public microarray datasets for me, and has been in charge of executing the biomarker-artifact overlap portion of this work. Many of the figures in this area have been generated in part by JT's code. JT was also my assistant while I did all of the PCR work discussed here.

Dr. R. Mitchell Parry recently joined our lab as a postdoctoral fellow and has often provided a fresh perspective on this project that has challenged me to make many of the improvements which are now the hallmarks of caCORRECT's success. Mitch has also helped streamline and document the caCORRECT code, saving me countless hours of computation. I would especially like to thank Mitch for lending me his knowledge on spectral unmixing and matrix factorization techniques which now permeate my work.

I would like to thank Dr. May Wang, my thesis advisor, for, among many other things, consistently convincing many very important people to listen to and devote resources to the ideas and investigations of her graduate students. As a corollary, I would like to thank the other members of my thesis committee for devoting time and resources to my ideas and investigations.

In addition to those specifically named here, I would like to thank the multitude of other students who have done important research with me, even if it did not end up in this document. I would also like to express my appreciation for my family and friends for their love, support, proofreading, and general ability to look interested when I talk about research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

caCORRECT	Chip Artifact Correction (a novel microarray processing system)
CC	Clear Cell RCC
CHR	Chromophobe RCC
FC	Fold Change
IHC	ImmunoHistoChemistry
MAQC	the MicroArray Quality Control Project
MAS	MicroArray Suite (a gene expression calculation)
MM	MisMatch Probe
NNMF	Non-Negative Matrix Factorization
ONC	Oncocytoma RCC
p	P-value (statistics)
PCA	Principal Component Analysis
(RT)-PCR	Polymerase Chain Reaction (for quantifying RNA expression)
PM	Perfect Match probe
QC	Quality Control
QD	Quantum Dot
RCC	Renal Cell Carcinoma
ROC	Receiver Operator Characteristic (for comparison of classification teqchniques)
RMA	Robust Multichip Average (a gene expression calculation)
SAM	Significance Analysis of Microarrays (a gene ranking method)
SVM	Support Vector Machines (a classification technique)

SUMMARY

Translational biomedical informatics is the application of computational methods to facilitate the translation of basic biomedical science to clinical relevance. An example of this is the multi-step process in which large-scale microarray-based discovery experiments are refined into reliable clinical diagnostic tests.

The quality of microarray data is a major issue that must be addressed before microarrays can reach their full potential as a clinical molecular profiling tool for personalized and predictive medicine. The FDA has completed phase-I of the MicroArray Quality Control (MAQC) project, and is currently developing guidelines and standards on microarray data reporting, quality control, and data analysis [1]. The current status of microarray quality control (QC) and noise reduction however, is still a controversial collection of tools and methods. While competing model-based tools such as dChip [2, 3], MAS5.0 [4], RMA [5, 6], and PLIER [7] have been developed to improve the quality of microarray gene expression data, these tools fall short in two important areas (1) they do not incorporate adequate spatial information into the outlier detection methods and (2) they do not incorporate outlier information into their normalization routines. The methodology discussed in this dissertation, called caCORRECT, addresses these deficiencies and seeks to replace or augment existing technologies in order to improve the translation of microarray data to clinical relevance [8-12].

As a case study to validate and demonstrate the usefulness of caCORRECT, the entire workflow of biomarker discovery was executed for the clinical problem of classifying Renal Cell Carcinoma (RCC) specimens into appropriate subtypes [13, 14].

Two biomarkers are discovered, NNMT and PRKAB1, which are able to separate the chromophobe and clear cell subtypes of RCC with perfect accuracy for all of the samples tested. To translate this discovery into a clinically relevant test, improvements are made to the reliability of quantum dot based immunohistochemistry [15, 16].

CHAPTER 1

INTRODUCTION

The central theme of this dissertation is the quality control of translational biomedical data, with a focus on gene microarrays and quantum dot immunohistochemistry, in an effort to improve reproducibility and reliability. To understand the potential impact of quality control, it is important to study it in the context of its use. This introduction aims to give the reader proper context, beginning with a definition of translational bioinformatics, followed by a description of the typical microarray experiment. We then discuss the current state of art in microarray quality control, followed by an overview of microarray biomarker selection, including the machine learning topics of classification and feature selection which are essential to developing high impact clinical solutions. Finally, we end with a discussion of an emerging frontier of translational biomedical informatics quality control—quantum dot based immunohistochemistry. Some of the work in this chapter is based upon my 2006 book chapter on microarray analysis [17], and my 2009 paper on quantum dot quality control [15].

Translational Bioinformatics

The ultimate goal of microarray analysis is the generation of reliable clinically relevant markers for disease and associated decision rules. This is a multi-step process that converts raw microarray data into biomarkers for clinical use. Proper and rigorous computational analysis is essential to the repeatability, reproducibility, and reliability of microarray data [18-35]. A schematic of the Translational Bioinformatics Pipeline is shown in Figure 1. First, raw microarray data are converted into useful gene expression data using any number of regression methods. Values of gene expression from multiple

biological samples are then used to select important features or to build predictive rules called classifiers. The results of feature selection and classification are lists of biomarkers that are appropriate for classifying the data into groups such as benign or malignant. These biomarker candidates are then validated through more specific measurement techniques such as PCR, and later Quantum Dot (QD) Immunohistochemistry (IHC). This dissertation aims to improve the reproducibility and reliability of the translational bioinformatics pipeline with novel contributions to the quality control of microarray data as well as QD-IHC.

As seen in the figure, microarray analysis is modular, and a small improvement in any step of the process has the potential to increase the performance of other steps downstream. Furthermore, the process is highly computational in nature due to the large volume of data (more than 200,000 probe measurements per chip) that must be analyzed. The translational nature of the pipeline is due to the increasing clinical relevance of results obtained at each step, concluding with clinically relevant, *in-vitro* diagnostic tests.

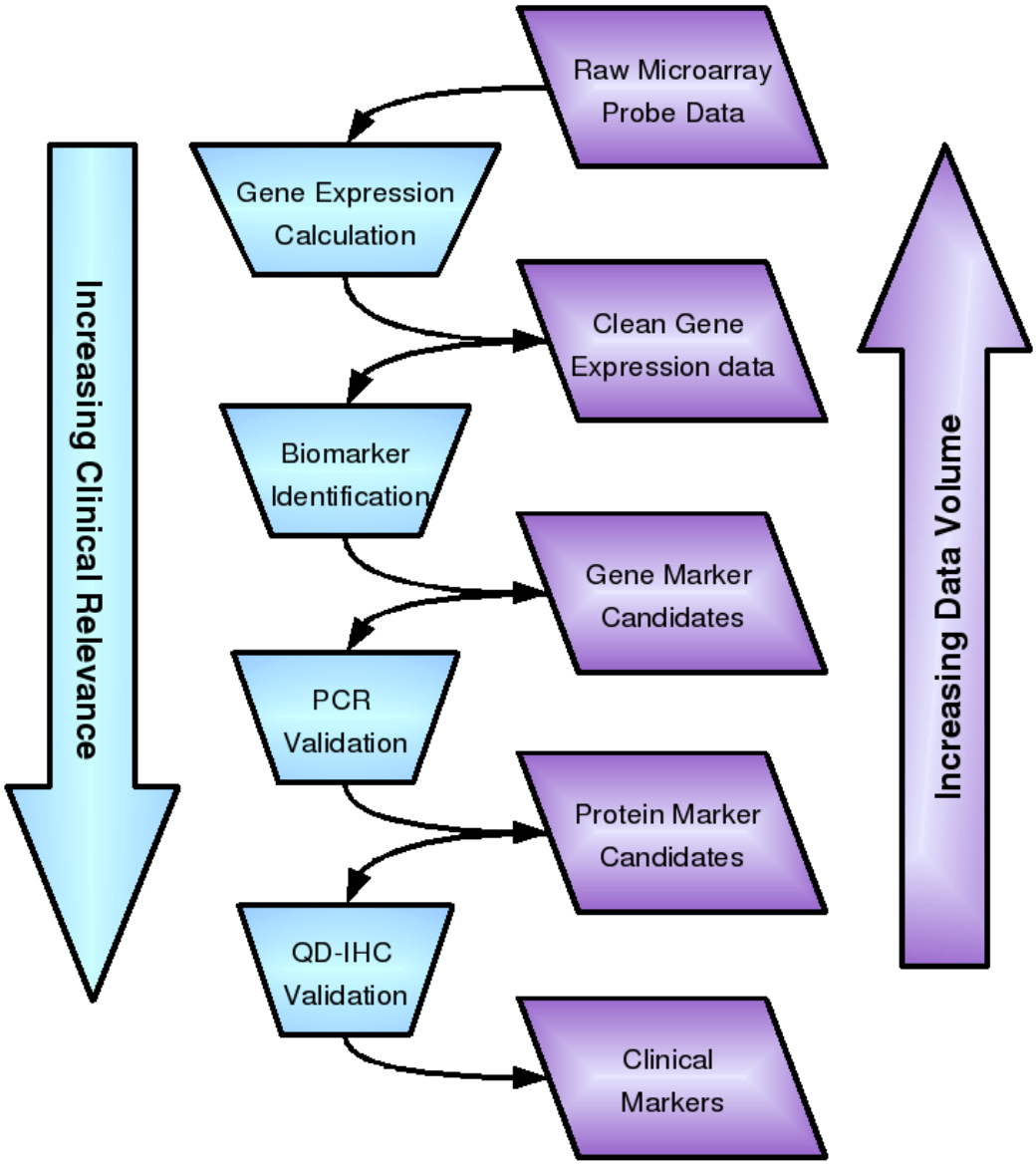

Figure 1: Translational Bioinformatics Pipeline.
Purple items represent data, while blue items represent data processing steps.

Gene Microarrays

Microarray analysis is a fast growing field, simultaneously harnessing advances in semiconductor manufacturing, molecular biology, medicine, and computation to provide an unprecedented genome-wide view of a biological sample. For the past ten years, microarrays have promised that the information from a single microarray might be used to tell a doctor if a patient has cancer, what type of cancer it is, what the prognosis is, and what drug to use to best fight the cancer. As the field has matured, however, serious questions have been raised as to the reproducibility and reliability of microarray based prediction. Although the FDA has endorsed the use of microarrays as a clinical profiling tool [1], others have disagreed [30], and a major drawback of microarray technology remains that it is less accurate and less reproducible than other RNA quantification procedures such as PCR [1, 25, 28, 36]. The main goal of this dissertation work is to help microarrays achieve their full potential as reliable clinical profiling tools by leveraging novel quality control methods to improve the reproducibility of results.

A DNA microarray is, generally speaking, a matrix of short oligonucleotide probes attached to a hard surface for the purpose of selectively hybridizing with unknown DNA in a solution. Some of the first microarrays used for genotyping, such as those constructed in 1995 by Schena et al. were printed on glass slides with custom-built high-speed arraying machines [37]. The design, selection, length, construction and attachment of these probes vary depending on experimental design, and can be very diverse depending on the application. Affymetrix, for example uses photolithography to attach sets of different twenty-five base pair probes in different locations on a single chip. See Figure 2 and Figure 3 for cartoons of Affymetrix microarray layouts. Other manufacturers, such as Affymetrix's leading competitor, Illumina, prefer to use longer cDNA probes attached to an array of beads on a bed of fiber optics. Evolving from Southern blotting, DNA microarrays are able to achieve much higher densities of

information-gained per sample-used than other techniques such as blotting or PCR. Microarrays achieve this density through the creative application of semiconductor manufacturing techniques, pioneered by companies such as Affymetrix in the early 1990's. Modern microarrays are manufactured by many different companies, and can be constructed on glass, plastic or silicon microchips.

In addition to the DNA microarray, protein microarrays have also been constructed in recent years using similar technology. Protein arrays consist of specific agents spotted to a surface that attempt to selectively hybridize with proteins in an unknown solution. Protein microarrays can be spotted with monoclonal antibodies to determine the concentration of protein in a sample, or functional proteins can be spotted to the array to determine protein-protein interactions. While protein microarrays differ from DNA microarrays in what they target, they serve similar and parallel roles in the type information they provide. For the purposes of this document, we will discuss microarray analysis in the context of DNA microarray analysis, especially those manufactured by the most popular vendor, Affymetrix, but the same concepts and strategies will apply reasonably well to any quantitative microarray platform with only slight modification. For an example of this, see our extension of caCORRECT to the Illumina platform [10].

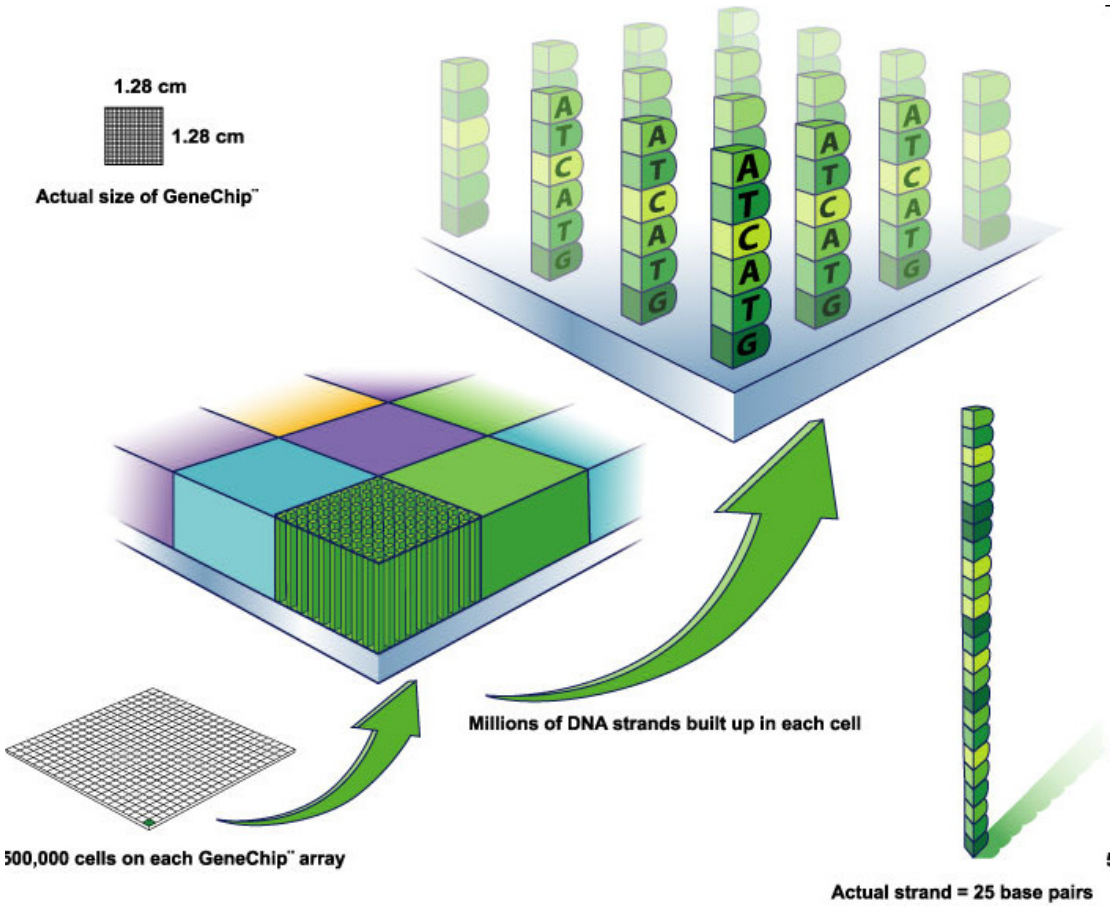

Figure 2: Cartoon Depicting a Single Feature on an Affymetrix GeneChip® Microarray. Image courtesy of Affymetrix.

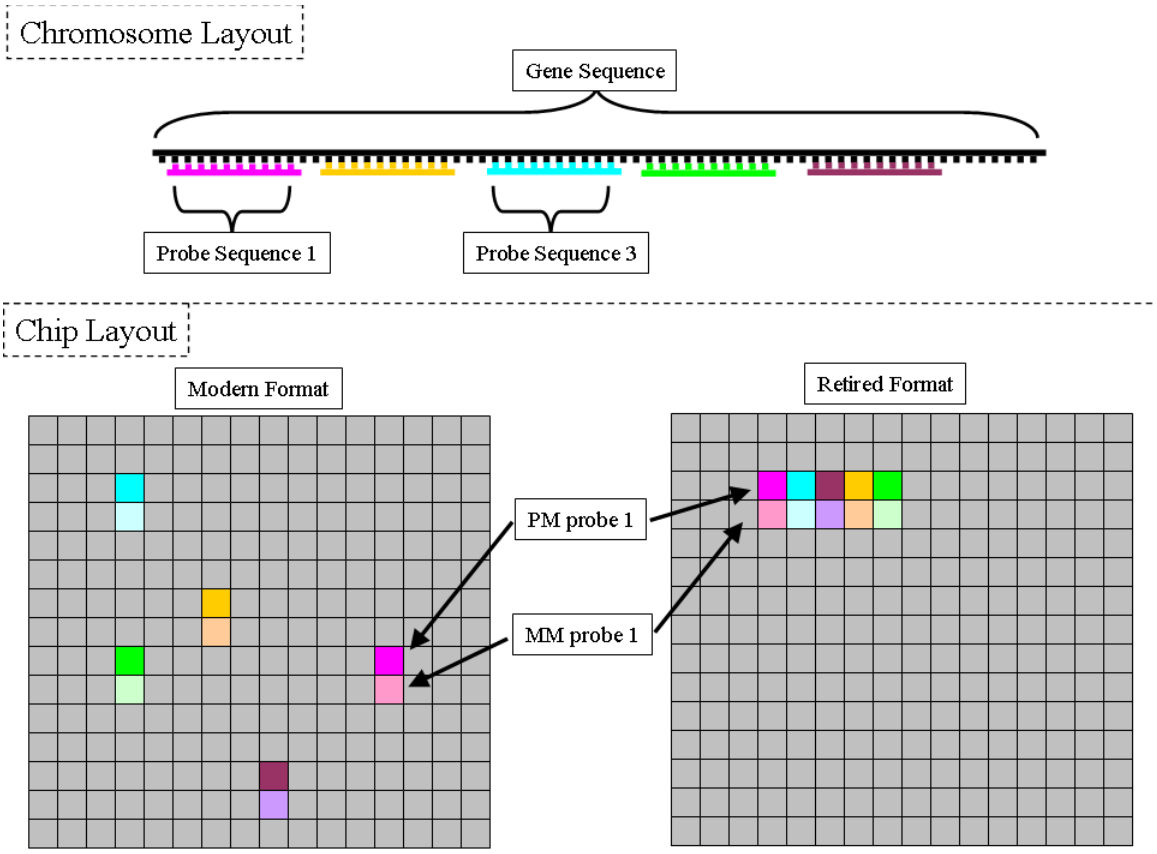

Figure 3: Cartoon Illustrating Microarray Probe Layout.
Square-toothed lines represent single stranded DNA. The top panel shows how multiple probes target a single gene sequence. The bottom panel shows how the probes of a single probe set are arranged on modern and older Affymetrix arrays. Note that the typical gene is measured by 20+ probes, and the typical array has more that 500x500 different probes printed on it. Perfect match (PM) and mismatch (MM) probes are shown in dark and light colors, respectively.

The basic procedure for using a DNA microarray is as follows. First, RNA is extracted from a biological sample. Next, the RNA is amplified to create fluorescently-labeled, complimentary cDNA, which is much more stable than RNA. The cDNA solution is then washed over the microarray and allowed to hybridize with the probes on the array (see Figure 4). The complimentary-base pairing of DNA is such that only cDNA that is a specific match for the probes on the array can attach properly. Even a single mismatch in a sequence will lead to suboptimal hybridization. In fact, Affymetrix arrays are designed with built in "mismatch" (MM) control sequences adjacent to each "perfect match" (PM) sequence to act as controls for non-specific binding. Unattached cDNA is then rinsed from the microarray, leaving only fluorescent cDNA that is complimentary to the probes on the microarray. A laser is then used to read the fluorescence levels at every spot on the array. These fluorescence levels are then quantified and collated with the probes they represent, and this "probe intensity" data is saved for further analysis.

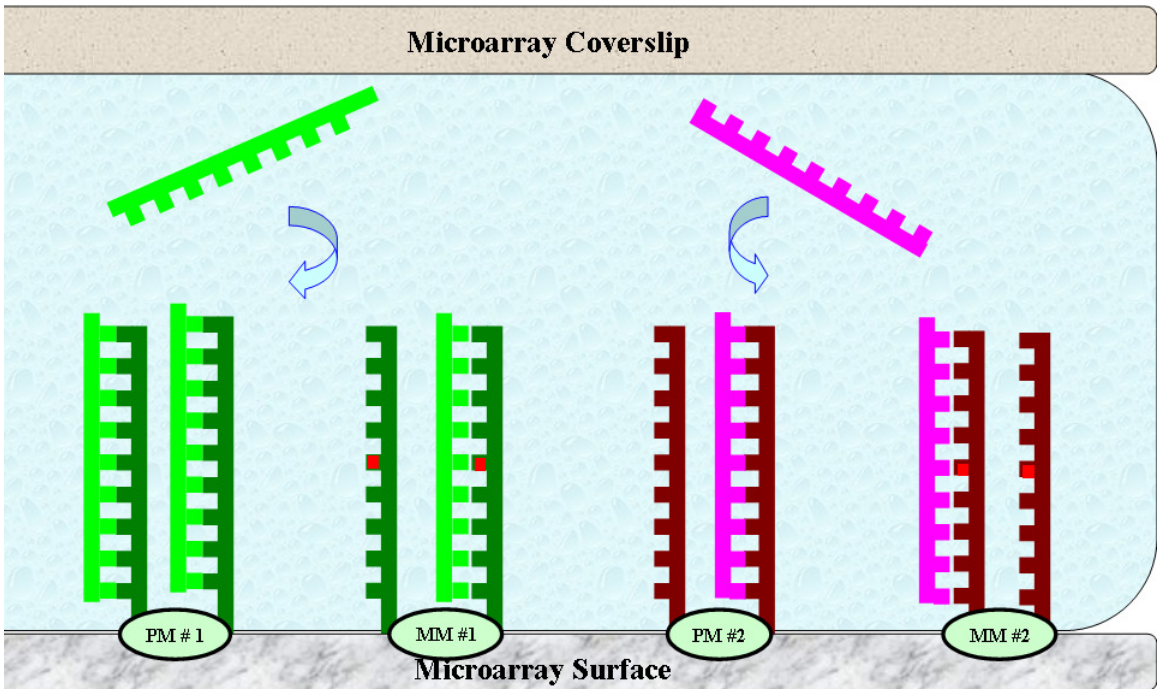

Figure 4: Cartoon Illustrating Microarray Probe Hybridization.
Square-toothed lines represent single stranded DNA. Perfect match (PM) and mismatch (MM) probes are shown in dark colors, while fluorescent cDNA from the sample being tested is shown in lighter shades. The two strands shown for each location are only indicative of the millions of probes on an actual microarray. MM pairs are indicated by a single non-complimentary base in the middle of the sequence, shown in bright red.

Microarray Experiments

Microarrays can be adapted to form a comparative study by using two samples simultaneously tagged with two differently colored fluorescent markers. These samples can then be assayed on the same chip and viewed with a dual-color scanner. Relative fluorescence values between colors are then recorded instead of absolute fluorescence for a single-color experiment. This work chooses to focus on single-color microarray data, but it would be fairly straightforward to adapt to dual-channel arrays as well.

The main advantage of the microarray is that it offers high-throughput genome-wide analysis by simultaneously measuring the expression levels of tens-of-thousands of different transcripts. Microarrays are capable of generating such vast amounts of data that they have the potential to eclipse the previous paradigm of single-target experimentation. Before microarrays, the prospect of a scientist beginning even a few hundred 'shot in the dark' gene expression experiments would have been ludicrous, but now with microarrays this can be performed relatively easily, yielding results in a matter of days.

One shortcoming of microarrays is that they offer primarily a tissue-wide view, meaning that single-cells cannot be viewed without significant RNA amplification that can alter relative expression levels of transcripts within the cell. Further complicating matters, standardization of RNA extraction and amplification is not perfect, and microarrays prepared in different laboratories often carry bias that clouds results [36]. In addition to variation in procedure, there is also wide variation in the microarrays themselves. Multiple platforms exist to choose from, with multiple companies offering different designs and libraries of transcripts. Some companies offer customized microarrays with user-specified transcripts, yet many researchers choose to print their own arrays. This variation in chip makeup makes it very difficult to compare results with other laboratories. To help combat standardization issues, a standard for Minimum Information About a Microarray Experiment (MIAME) was proposed in 1999, and is still

being improved today. More information on MAIME, and the current MAIME standard checklist can be found at http://www.mged.org/Workgroups/MIAME/miame.html [38]. Manufacturing inconsistencies such as those that arise from the tips used to print the microarrays, or inconsistencies in the washing and drying procedures, also contribute to noise that can seriously hamper results. To correct for chip noise, a variety of statistical and quality control methods have been proposed, and are discussed later [5, 6, 12, 25, 39-50].

The microarray is a diverse platform, and can be used in a variety of experiments. Some of the earliest experiments to use microarrays studied gene expression during the life cycle of yeast [51]. Modern experimentation, however, has expanded to comparative studies such as healthy versus cancerous tissues, or time-series analysis studying the progression of disease. While applications may vary, the most common task for a microarray experiment is to identify a small set of "biomarker" genes that can differentiate between two or more varieties of biological sample. These interesting genes can then be used as targets for more specific assays to be used as screening tests, or the genes may be studied further to learn more about the mechanisms governing the problem at hand.

As with any experiment, repetition and balance are essential to successful results. This may be especially important with microarrays, because of their sensitivity to manufacturing defects or errors in preparation. Unfortunately, due to the relatively high cost of microarrays, many researchers cannot afford such extensive repetition and cross-validation. Fortunately, depending on the problem being investigated, different experimental setups may be used to maximize the resources available. The problem of optimal design is especially relevant when multiple treatments and two-channel arrays are used. While the first instinct might be to use a procedure where all samples are compared to one reference, it has been shown that when larger ($n>5$) numbers of treatments are studied, loop and modified loop designs perform better than the standard

reference design. Furthermore, while it is clear that using more microarrays for an experiment is better, there are exceptional gains in marginal efficiency when two more (n+2) or twice as many (2*n) arrays are used [52]. In the end, however, even the best of experimental designs can be crippled by laboratory or manufacturing error, resulting in noisy or defective data and poor results. This effect can be alleviated with proper quality control and assessment procedures like the ones proposed in this document.

Microarray Quality Control

The quality of microarray data is a major issue that must be addressed before microarrays can reach their full potential as a clinical molecular profiling tool for personalized and predictive medicine. The FDA has completed phase-I of the MicroArray Quality Control (MAQC) project, and is currently developing guidelines and standards on microarray data reporting, quality control, and data analysis [1]. The current status of microarray quality control (QC) and noise reduction however, is still a controversial collection of tools and methods.

Much work has already been done to try to improve the accuracy of derived gene expression data from microarray chips. Efforts can be generally categorized into two strategies: (1) array Quality Assessment (QA) methods, and (2) robust gene expression calculations. The comprehensive system proposed in this dissertation, caCORRECT, uniquely falls into both categories. Table 1 shows a comparison of features for caCORRECT and popular QA methods, whereas Table 2 shows a comparison of features for caCORRECT and popular gene expression calculation methods.

In one of the earliest attempts at array quality assurance, Yang et al. proposed a model based method for flagging statistical “weak spots” on dual-color spotted arrays [44]. Yang’s method’s main goal was to eliminate nonsensical ratio measurements made from probes with very low intensity values. Yang et al.’s method is worth mentioning in particular, because it is the closest published example to caCORRECT’s artifact aware

normalization scheme. Yang et al. describe a way to estimate a scaling factor for use on the different-colored intensities of two-color arrays by ignoring "weak" probes. Here, they make the assumption that all of their "weak" probes are manufacturing errors and thus should not factor into the scale factor calculation. While similar in concept to caCORRECT's scheme, it is rather trivial in comparison, and relies on human selection of threshold values for "weak" spot detection. Furthermore, the normalization for Yang's method is completely isolated from outlier detection, whereas caCORRECT's normalization is allowed to iteratively improve the sensitivity of outlier detection.

A serious flaw in all of the most popular gene expression calculations is that they do not incorporate spatial chip layout information into their outlier identification schemes. One QA method that does take spatial effects into account is that of Reimers and Weinstein, which was later named SmudgeMiner [40]. This system can be used to visualize regional biases across high-density chips. Citing factors such as temperature, liquid flow rate, RNA diffusion rate, and edge effect, they showed that significant regional biases are common. In addition to localized background calculation, SmudgeMiner produces a comprehensive quality score for each chip by measuring the correlation of each probe's expression level to that of its neighbors. While this application may provide quality score information, it does not allow correction of these artifacts. Users are then faced with a difficult choice: abandon a chip, or proceed knowing that artifacts exist. Another system, arrayMagic [48] has the same drawbacks as SmudgeMiner i.e. merely assessing quality without providing an avenue of improvement beyond discarding whole chips.

In the development of new methods for microarray QA, Brodsky et al. proposed a novel method of using clustering of gene expression profiles across microarrays to indicate quality [53]. First, gene expression profiles are clustered, and then the randomness of the clusters' distributions across the microarrays are measured. Second, the spatial distributions of high and low expressed genes are monitored on each sample

for randomness. These two qualities are then used to identify artifactual genes. While Brodsky's method does include intelligent use of chip layouts for artifact detection, it falls short as a major contributor to modern translational bioinformatics by not supporting Affymetrix arrays. Furthermore, its methodology for dealing with artifactual data (a local mean replacement) is trivial, and of debatable use in comparison to caCORRECT's more informed model-based imputation. Furthermore, neither the method of Brodsky et al. nor any of the other quality assessment methods surveyed here have been applied to the Illumina platform, although caCORRECT has [10].

The best attempt to date (other than caCORRECT) at spatial outlier detection for Affymetrix arrays is a system called Harshlighting [41, 50], which was published after my first attempt at spatial outlier detection [8, 9], after the launch of what is now the caCORRECT website (at the time, called ChipQC, and later PADRE), but before journal publication of the definitive caCORRECT paper [11]. Harshlighting is similar to caCORRECT in that it leverages image processing techniques such as sliding windows and background assessment to identify an assortment of compact and diffuse artifacts. Harshlighting, however, performs image processing on an error image which is a simple distance from the median. Not only does this formulation ignore the natural variance of probes, but it also neglects to account for global chip to chip variation which may lead to whole chips being discarded that may be correctable with a simple normalization step. Furthermore, the authors of Harshlighting point out the appearance of "ghosting" artifacts i.e. the incorrect appearance of artifacts on clean chips as a result of comparison to a severe artifact on a different chip in a batch. Whereas Harshlighting attempts to correct for this phenomenon by using a median in its error heat map calculation (as opposed to the more outlier-sensitive measure, the arithmetic mean), caCORRECT completely avoids the problem by iteratively identifying artifacts and omitting them from calculations altogether (see equations 2.1-2.4 for details).

In summary, highlighted by Table 1, caCORRECT represents an improvement over existing QA methods by (1) global normalization before outlier identification, (2) model-based outlier replacement, (3) iteration of outlier removal and normalization, and (4) application to the two leading chip manufacturers, Affymetrix and Illumina.

The second front of microarray quality control is robust gene expression calculation. Affymetrix microarrays, for example can be processed with Affymetrix's own Microarray Suite (MAS5.0) [4], GeneChip Operating Software (GCOS), or Probe Logarithmic Error Intensity Estimate (PLIER) [7], but alternatives such as dChip [39], PerfectMatch [54], and RMA [5, 6] also exist. These programs include good quality control measures such as normalization, background correction, and robust model fitting in an attempt to determine gene expression from multiple probe values. Many of them provide a visualization feature showing where outlier probes are located on the chips, but yet they (1) do not include this spatial information in their outlier detection, and (2) they do not incorporate outlier information into their probe normalizations (see Table 2). This is troubling, considering the existing body of knowledge on spatial chip artifacts, including one study which showed that gene co-expression results were found to be correlated with the proximity of genes on the microarray chip, an otherwise randomized event [56]. A gallery of such spatial artifacts (derived from caCORRECT artifact heat maps) can be found at the ArrayWiki website: arraywiki.bme.gatech.edu [57] or at the RMA express image hall of fame page at plmimagegallery.bmbolstad.com [58].

dChip includes methods for statistical detection of array, probe, and spot artifacts, but does not include spatial information in its algorithms, nor does it incorporate outlier information into its normalization scheme. RMA as well lacks these innovations which are provided by caCORRECT. Furthermore, the global normalization scheme employed by dChip, which selects a single chip as a template for normalization, is especially prone to the pitfalls of normalization in the presence of artifacts, just as RMA's quantile normalization, which is later discussed in chapter 2, Figure 10.

While the popular tools dChip and RMA suffer from warping in the presence of artifacts due to their chosen normalization techniques, MAS5.0 skirts the problem by not normalizing probes at all. Accordingly, MAS5.0 has been shown to be less accurate than dChip or RMA [5, 49, 50, 59], but it remains popular for clinical applications where its ability to process one chip at a time is appealing.

Table 2 shows areas of novel protocol improvements with respect to existing gene expression calculation software. We will later support the claim that these innovations are indeed improvements in chapter 3 as we give evidence that shows caCORRECT improves the accuracy of gene expression and the reliability of biomarker discovery over RMA and MAS5.0. We chose to focus on these two methods due to their overwhelming popularity among FDA MAQC phase II participants (data not shown).

Table 1: Comparison of caCORRECT to other Quality Assurance Methods

	caCORRECT [11]	SmudgeMiner [40]	Brodsky et al., 2004 [53]	ArrayMagic [48]	Harshlighting [41, 50]
Spatial information used for artifact detection	Yes	No	Yes	No	Yes
Global normalized	Yes	No	No	No	No
Treatment of artifact data	Model Imputation	N/A	Neighborhood Mean Replacement	N/A	Probe Median Replacement
Iteration of normalization and outlier detection	Yes	N/A	No	N/A	N/A
Support for Affymetrix	Yes	Yes	No	No	Yes
Support for Illumina	Yes	No	No	No	No

Table 2: Comparison of caCORRECT to other Gene Expression Calculation Software

	caCORRECT [11]	**dChip [60]**	**RMA [5, 6]**	**MAS5.0 [4]**
Visualization of errors	**Yes**	**Yes**	**Yes**	**No**
Spatial information used for artifact detection	**Yes**	**No**	**No**	**No**
Probe normalization	Quantile	Invariant set fit to template chip	Quantile & Background Correction	**No**
Artifact-aware probe normalization	**Yes**	**No**	**No**	N/A
Iteration of normalization and outlier detection	**Yes**	**No**	**No**	N/A
Gene expression calculation given for all genes	**Yes**	**Yes**	**Yes**	**No**

Microarray Analysis and Biomarkers

Microarray analysis is, essentially, a pattern recognition problem with two goals. The first goal is classification—'given a sample of unknown type, what class does the sample belong to?' The second goal is the more fundamental goal of feature selection—'which features carry the most information to help us in the problem of classification?' These two goals of classification and feature selection are not independent, and just as the success of classification relies on good features, feature selection may receive feedback from the success of classifiers.

In the language of microarray analysis, features refer to genes-- expression levels of RNA as measured by a microarray. A single observation of the feature vector is the sample, which is RNA from a biological source, such as a tissue biopsy measured via a single microarray. Multiple samples comprise the many data points of the dataset, which is often divided into training and testing subsets for further analysis. In the context of microarrays, classes can include normal or diseased tissue, malignant or benign cancers, or even subclasses of cancers which may aid physicians in selecting appropriate treatment.

Classification

Sample classification can be done in an unsupervised or supervised manner. Unsupervised methods make no assumptions about the correct class of a sample, but attempt to group samples into classes based on their similarity or distances from one-another. Supervised methods, on the other hand, attempt to learn a decision rule from a training set of labeled data, and then use that rule to classify unknown samples in the test set. While supervised methods generally produce more powerful classifiers, unsupervised methods can be useful in discovering previously unidentified subtypes. These techniques are not mutually exclusive, and clustering (an unsupervised method), for example can be

helpful as a first step before further analysis or as an intermediate step in more powerful supervised techniques such as SAM or SVM.

While a survey appears below, more general discussions of modern pattern classification methods are widely available in texts such as *Pattern Classification* by Duda Hart and Stork [61].

Hierarchical Clustering

One of the simplest and most common forms of unsupervised classification is hierarchical clustering. Hierarchical clustering is performed by iteratively grouping the two closest samples as determined by a distance metric. Euclidian distance and correlation are two commonly used distance metrics, but many others exist. The iterative grouping of samples continues until one of two conditions is met: 1) the number of clusters equals the desired number of class separations, or 2) the distance between clusters reaches some threshold. The results of hierarchical clustering can then be incorporated into a simple prediction rule: an unknown sample is assigned to the class which it is nearest to. While hierarchical clustering is fast and simple, it becomes decreasingly useful with smaller sample sizes, or poorly separated classes.

K-means Clustering

Another unsupervised classification commonly used is k-means clustering. K-means clustering attempts to minimize the distortion of each class by iteratively recomputing class centers using an Expectation-Maximization approach. K-means clustering has the advantage of allowing the user to specify the number of clusters that they are looking for, and the advantage of finding optimal or near-optimal clusters. This generally leads to better results for classification than a hierarchical cluster. Specifying the number of classes could be a disadvantage, however if the goal of clustering is to identify new subclasses of disease.

SAM/PAM

Significance Analysis of Microarrays (SAM) and Prediction Analysis of Microarrays (PAM) are two recently proposed methods of feature selection designed specifically for microarrays. SAM, developed by Tusher, Tibshirani and Chu, generates a SAM test statistic for each gene [62]. The SAM statistic is like a t-statistic, but uses a small positive constant added to the standard deviation in the denominator. This addition artificially increases the variance so that the test is less likely to pick up genes with very low, hard to duplicate expression levels as significant. The significance of this SAM test statistic is calculated with a permutation test, which tells how rare the calculated test statistic is among simulated statistics calculated from randomly relabeled samples. SAM then provides a ranking of the genes by ordering the significance of the SAM statistic. While SAM and other statistical methods, such as fold change or t-test p-value thresholding, are good at finding differentially expressed genes, they can discard potentially important genes with more nonlinear, yet significant characteristics.

The other method created by the Tibshirani group, PAM, uses soft thresholding to "shrink" the list of genes until only a core of useful genes remains [63]. This shrinking process is accomplished by eliminating genes from the classifier if the gene's expression centroid is similar for all classes. The resultant shrunken list of genes then have more distinct centroids than the original gene list, and are thus better-suited features for classification. To classify an unknown test sample using PAM, only the shrunken list of genes are used as features. The test sample is then placed into the class where the distance between the test sample and class centroid is minimized. A correction factor may also be added to account for the *a priori* probability of class membership. PAM, like SAM, also results in a ranking of significant genes, which may be good features choices for a classifier.

Support Vector Machines

A support vector machine (SVM) is a classifier in which a maximal margin hyperplane is generated that separates multiple classes of data points[64]. The traditional form of the SVM problem handles only two classes, but may be extended to more than two classes. Data points can be multidimensional without greatly affecting processing performance of the algorithm, which makes the SVM well suited for microarray data points, which may consist of thousands of features. Furthermore, SVM generalizes well even for small sample sizes, which is usually the case with microarray experiments. A major advantage of SVM over other methods is that through the use of a kernel function, the SVM is able to make use of linear and nonlinear trends in the data. The nonlinear capabilities of SVM are especially showcased when analyzing multiple features. In this case, subtle biochemical interactions such as inhibition, activation, and redundant function can be seen and used to create complex, nonlinear classification rules. In addition to being used for classification, the SVM can also be used effectively for wrapper-based feature selection as discussed later in the "Feature Selection" section. The versatility and scalability of the SVM combine to make it an attractive and popular choice for microarray analysis.

Feature Selection

Feature selection in the context of microarray analysis is synonymous with finding a set of differentially expressed genes. Because these genes will ultimately be used as input to a classifier, feature selection can also be thought of in the context of supervised analysis. If each gene is used independently to build a classifier, it follows that the genes which produce classifiers with lowest error in prediction rates contain the most discriminating information, and are thus good candidates to serve as biomarkers. This is the basis of *filter* methods of feature selection. Alternatively, the best performing set of features is not necessarily composed of the set of the best individually-performing

features. Finding such a best performing set of genes is the basis of *wrapper* methods of feature selection.

For filter methods, features are selected or ranked individually based on their independent discriminatory power. Common methods for ranking include the *p*-value from a t-test, the observed fold change of gene expression between classes, or the estimated error of a single-feature classifier using just that gene. After ranking genes, a binary selection criteria, such as "top 50 genes" or "all genes with fold change greater than 2" is applied to create a list of features on which to build a predictive model. Filter methods are considered suboptimal in that they do not account for interactions among features. This is an especially big problem in the context of gene microarray data, which are known to be highly co-regulated, highly networked, and thus not independent. Using filter methods may, therefore, lead to selection of redundant features which add complexity without adding discriminating power. For this reason, wrapper methods are sometimes preferred.

For wrapper methods, features are evaluated in sets for their combined ability to contribute to a successful classifier. To begin, a classification method and an error estimation technique are selected, and then used to rank perspective feature sets based on the error estimate of classification. The typical microarray has on the order of thousands of genes, which makes an exhaustive search of even pair wise feature sets a daunting task. Exhaustive searches of sets larger than 3 or 4 are usually impossible to complete in a reasonable time. To counter the factorial growth of exhaustive searches, other optimization routines, such as iterative feature addition/subtraction or genetic algorithms may be employed [65].

Error estimation

When assessing the performance of any classifier, it is important to know how reliable the classifier is, or, more precisely, how accurately will the classifier perform

given new, unknown input? At the root of this question lies the selection of an error estimation method—the method used to estimate what the error rate of the classifier *would truly be* in a real application. While an improperly low error estimate during feature selection will lead to false-positive results, and can lead to fruitless and costly validation studies, an improperly high error estimate will lead to false-negative results, and may exclude a useful feature from selection. In the context of microarray analysis, false-positive results are undesirable and costly but may be corrected via proper validation. False-negative results, on the other hand, cannot be recovered from because they will not be investigated any further. Another important property of error estimation is that it is often integral in the design of the classifier itself, because it can be a necessary parameter for feature selection, selecting modeling parameters, or even convergence of the classifier. Fortunately, many different strategies of error estimation exist, such as resubstitution, cross-validation, and bootstrapping [35, 66].

The selection of an appropriate error estimator involves a delicate balance of computational effort, bias, and variance. Some methods, such as resubstitution provide under-estimates of the true error, while others such as bootstrapping are often biased towards high estimates of the true error of classification. Besides bias, variance is also an important characteristic to consider when choosing an appropriate error estimator. Even if the error estimate is unbiased, a high variance will undermine results by often estimating an error that is far from the true error. Finally, computational cost must also be considered when selecting an error rate estimator. Often the best, most accurate error estimation methods involve repetitive sampling of training and testing datasets for each feature set being analyzed. With each iteration, a new classifier must be built, which will slow down the overall progress of the algorithm. Methods such as resubstitution, however offer minimum computational effort because they require only one classifier to be built for each feature set. For problems with large search spaces, such as finding an optimal combination of biomarkers, the computational costs of error estimation have an even

larger impact, and the added accuracy of good estimation may be eclipsed by the practicality of simpler methods.

Resubstitution

Perhaps the simplest method of error estimation and one of the most computationally inexpensive methods is resubstitution. Resubstitution is by definition the use of the entire dataset as both the training and testing dataset. This is equivalent to a student receiving an exact copy of the final exam to study from before he actually takes the test. In the same way that the student would likely score higher on the exam than he would have on random questions, the results of a resubstitution estimated error will be generally optimistic. As a corollary, the more complex or prone to over-fitting that the classifier is the worse that the resubstitution will perform by allowing the over fitting to proceed unchecked. It is usually a bad idea to use resubstitution to estimate classifier error, especially when the ratio of features to samples is low.

Cross Validation

As an alternative to resubstitution, cross validation attempts to correct for bias by separating testing and training sets. A k-fold cross-validation consists of dividing the data into k subsets, assigning one subset to be the testing set, and the remaining k-1 subsets as the training set. Cross validation can be run in either a *complete* or *iterative* manner. A *complete* cross validation executes exactly k times for a k-fold cross validation. In other words, once it divides up the data, it systematically leaves each subset out of the training set until each subset has taken its turn as the test set exactly once. A special case of complete cross validation is the 'complete-leave-one-out', or 'n-fold cross validation'. In this case, each single data point is left out exactly once, and the result is a nearly unbiased estimate of how the feature set might perform given brand-new data. With complete cross-validation, the number of classifiers that must be built for each error estimation is

capped at k. This is not the case for an *iterative* cross validation, in which many different partitions of the data into training and testing datasets are made. Iterative cross validations can, for example perform 100s or 1000s of independent segregations of the data, each time choosing a different set of data to act as the testing subset and the remainder to act as training data.

Bootstrapping

Another error estimation method similar to cross-validation is bootstrapping. Bootstrapping involves selection, with replacement of n samples from a dataset of size n. The classifier is built on the chosen n samples (including some duplicates), and the error is estimated on those samples which were not selected for training. Like cross-validation, this bootstrap procedure may be performed iteratively in order to increase the reliability of the error estimate. Also like cross-validation, bootstrap estimates tend to be slightly pessimistic estimates of classifiers built using all of the data, because the bootstrap classifiers are trained on smaller subsets of the data, and classification accuracy generally improves with sample size. To offset the pessimistic nature of bootstrapping, bootstrapping can be combined with the optimistic result of resubstitution in the ratio 63.2% bootstrap error to 36.8% resubstitution error to give the "632+ bootstrap" error estimation[67]. Because of this combination strategy, the 632+ bootstrap has the desirable property of being neither optimistic nor pessimistic and generally outperforms other methods in terms of both bias and variance [68]. Unless otherwise stated, all error estimates used in this dissertation are 632+ bootstrap estimations.

Information Leakage

For any type of feature selection method, care must be taken to separate the feature selection step from the error estimation step [23, 34, 69]. In other words, any samples which are used for estimating the error of a classifier must not have been used

either in the training of the classifier or in the selection of features. Any such information leak will lead to an overly optimistic estimate of future prediction performance. While this rule implicitly prohibits use of simple resubstitution as a method of error estimation, the 632+ bootstrap method is a notable exception to the rule. Separation of feature selection from training and testing are a major theme of the FDA MAQC-phase II work, which I have contributed to, and which is currently under journal review. All work in this dissertation strictly adheres to the above-described policies of "honest" cross-validation.

Quantum Dot Immunohistochemistry

Biomarkers can be identified, either by applying the previously mentioned machine learning techniques to high-throughput discovery technologies such as the microarray, or by searching the existing literature for well-established markers, such as ER, PR and HER2/neu for breast cancer. HER2/neu in particular, represents a good example of clinically-relevant molecular classification of tumors. Patients who overexpress HER2/neu can been treated with Trastuzumab, a recombinant monoclonal antibody against HER2/neu, in order to increase life expectancy [70]. Protein expression for this type of analysis is typically measured semi-quantitatively by a pathologist manually scoring the results of immunohistochemical (IHC) staining for the protein of interest.

To achieve a clinically relevant molecular profile for an unknown subtype of a less well studied disease, such as Renal Cell Carcinoma (RCC), a more reliable, and thus quantitative, technique than traditional IHC is required. Because of the complexity of RCC, and the multiple subtypes to be separated, is expected that a single marker will not be sufficient for accurate classification. Instead, it will be necessary to monitor the expression of a small panel of markers to build a proper classification rule. Because of this constraint, methods which can measure multiple markers simultaneously will be favored. Thus, traditional IHC, which measures only one protein's expression at a time,

will not be sufficient for building an accurate classifier, given the limited supply of tissue sample in an actual clinical setting. Protein expression in this study will instead be measured in clinical samples by staining with multiplexed Quantum Dot (QD) panels. QDs have been chosen as the reporter agent due to their ability to be multiplexed and their potential to be quantified [71-75].

Even with these promising pioneering studies, issues still remain that hold back the use of multiplexed QDs in a clinical setting as a quantitative tool. Much like the case with microarrays, there is a need for a similar application of quality control and computation to improve the reliability and repeatability of QD-IHC to the point where it may be used in a clinical setting. These issues are discussed here, and in chapter 5.

Multiplexed QD-IHC assays have been developed in a number of labs using direct primary antibody-QD (Ab-QD) conjugation [75-77]. Such studies are questionably quantitative, due to lack of large-scale reproducibility of custom Ab-QD conjugation, making translation to a clinical setting doubtful. A more translational protocol would instead use widely available commercial QDs, commercial antibodies, and simple chemistry. One such alternative approach, using primary antibody staining and biotin-streptavidin coupling to QDs in solution, has been achieved, but such protocols are not yet suitable for more than 2 markers [78]. Yet another approach involves use of a primary antibody stain followed by application of commercial QD-secondary antibody conjugates [75]. While this protocol is a promising candidate for clinical translation, multiplexing beyond 2 QDs relies on further optimization to reduce antibody cross-reactivity [74].

Quantification of protein using QDs relies on the assumption that the QDs actually attach to their intended targets. The basic scheme for a 4-plex QD stain using commercially available reagents is Figure 5. In this scheme, four different QDs are intended for four different protein targets. Problems arise with this protocol whenever primary antibodies added in step 1 are not completely saturated by the secondary-QD conjugates added in step 2. In such a situation, secondary-QD conjugates added in step 4

may errantly bind to primary antibodies from step 1 as well as their intended targets from step 3. Figure 7 shows a cartoon of this effect. This antibody cross-reactivity has been identified previously in literature as a serious obstacle to QD-IHC becoming useful [74]. Unsaturated primary antibody could occur due to any of: (1) Insufficient concentrations of QD-secondary conjugates in step 2, (2) Insufficient incubation time in step 2 to cause saturation, (3) Disassociation of primary and secondary antibodies during the subsequent washing and staining procedures between step 2 and 4, or (4) Other forces i.e. QD-antibody-valency effects which may cause one QD-secondary to displace another.

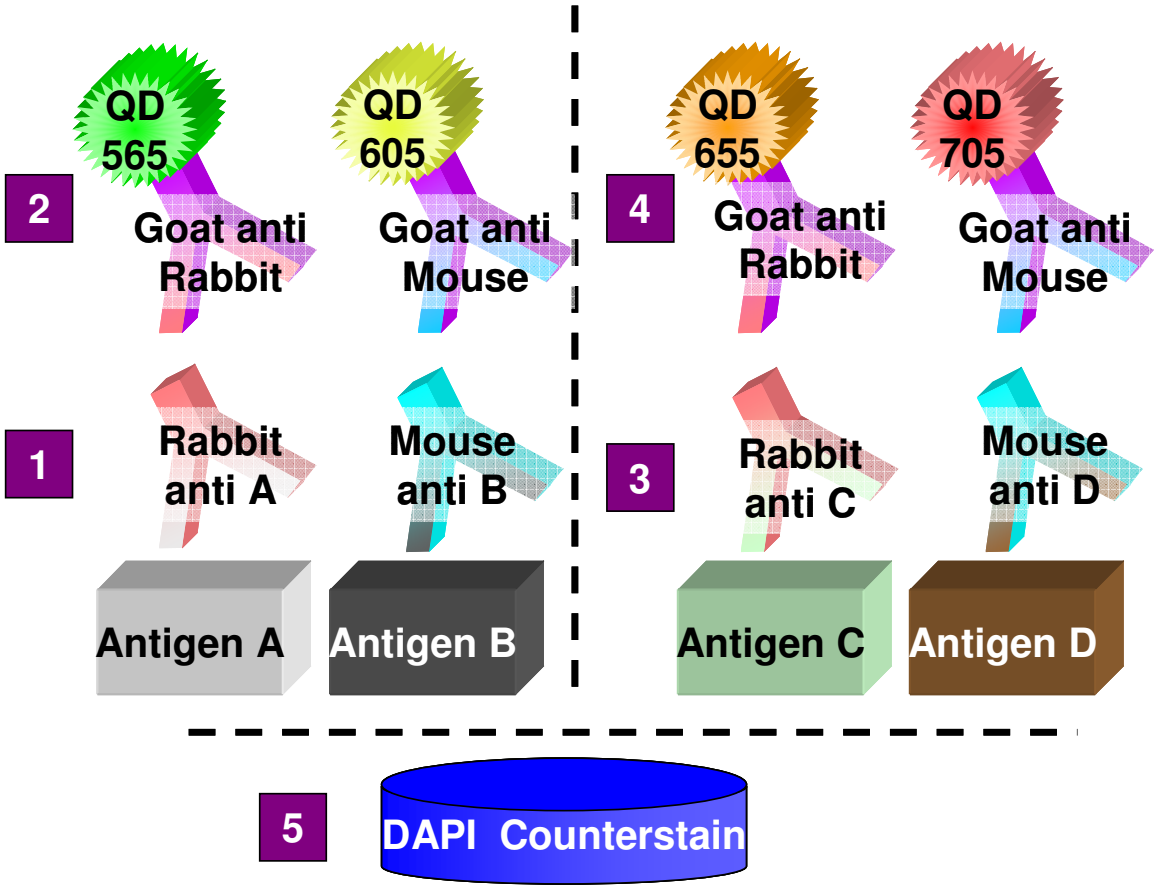

Figure 5: Multiplexed Immunostaining Workflow.
Step 1 and 3 are addition of primary antibodies, Steps 2 and 4 are the addition of QD conjugated secondary antibodies and step 5 is nuclear counterstain with DAPI. A 4-plex procedure is depicted, but Steps 3 and 4 may be repeated with different targets and QDs to increase the multiplicity of staining. This figure is reproduced from [15].

If the scenario depicted in Figure 6 were to occur, it would mean the presumption of antigen C in tissue locations which express antigen A. Importantly, this effect does not work in reverse. For example, QD565 does not bind to antigen C because no primary rabbit anti-C antibodies are present when QD565-anti-rabbit is introduced. First and foremost, such antibody cross reactivity should be avoided on a chemical and procedural level by minimizing the four previously mentioned pitfalls. In the absence of perfect staining protocols, software solutions, such as those outlined in chapter 5, are an attractive alternative for achieving reliability [15].

Unmixing of QD signals from tissue autofluorescence is another problem which leads to low sensitivity and quantitative accuracy in QD-IHC analysis of tissue sections [74, 76]. Some progress has been made to increase the signal to background ratio using the unique properties of QD i.e. long emission half-life [79], or photostabillity [75, 80]. Separation of QD from autofluorescent background has been claimed [81, 82], but the evidence supporting these claims is minimal, and methods to achieve these results remain a trade secret. Our own results discussed in Chapter 4 suggest that autofluorescence is a still major source of error in QD-IHC, and a definitive solution to the problem is yet to be published.

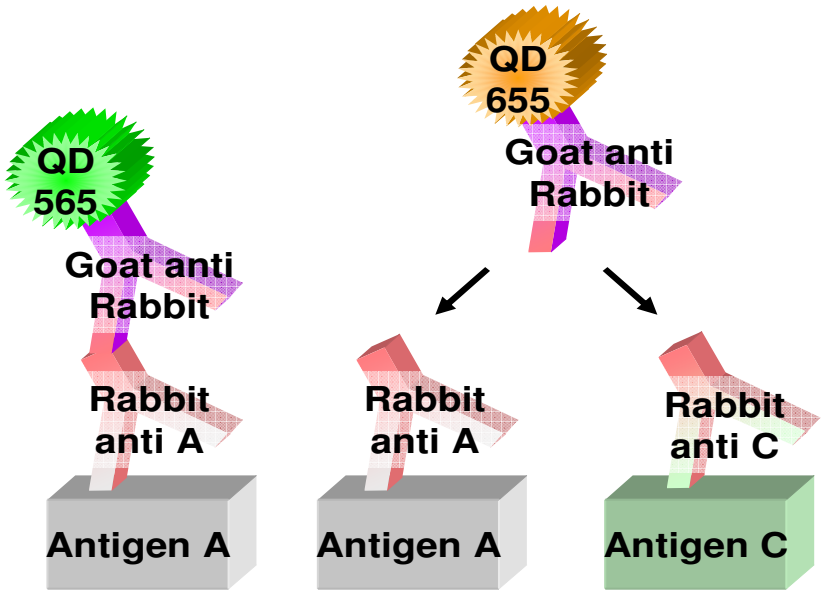

Figure 6: Undesirable Signal Crosstalk.
This cartoon illustrates a scenario where QD655 signal will be present at antigen A as well as antigen C. This occurs if there is unsaturated or unblocked primary anti-A antibody sites left over from the previous round of staining. This figure is reproduced from [15].

Structure of Dissertation

Motivated by the existing questions behind the reproducibility and reliability of microarray and QD-IHC data, as well as by the need for developing more reliable clinical tests, I have divided this dissertation into three specific aims:

Specific Aim 1: To develop a comprehensive microarray quality control algorithm to minimize the effect of common sources of noise on gene expression calculation

Specific Aim 2: To investigate the effect of microarray noise and quality control on the reproducibility of gene biomarker discovery

Specific Aim 3: To discover and validate novel clinical gene and protein biomarkers for cancer diagnostics.

Figure 7 summarizes the key areas where this dissertation improves the workflow for translating microarray-based biomarker discovery. As part of the proposed methodology for Specific Aim 1, algorithms were developed for microarray Chip Artifact Detection, Artifact Aware Probe Normalization, and Artifact Aware Expression calculation. These all serve to increase the reliability of gene expression, which was investigated as part of Specific Aim 2. To achieve Specific Aim 3, the Translational Bioinformatics Pipeline was executed for a case study in Renal Cell Carcinoma (RCC). The final result of this work represents a step towards quantitative, biomarker-based screening method for RCC subtyping, but is easily extendable to other molecular diseases.

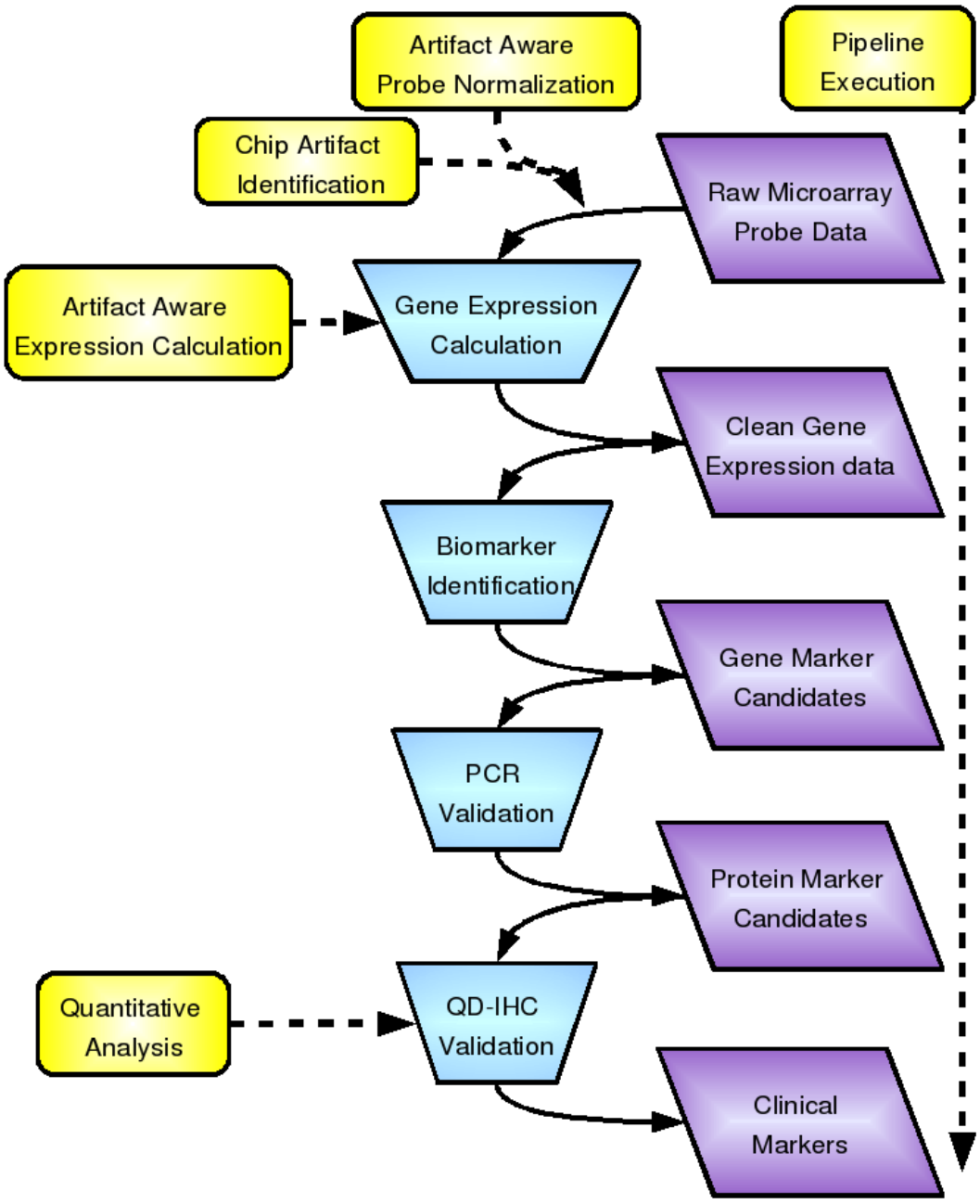

Figure 7: Expected Contribution to the Translational Bioinformatics Workflow. Purple items represent data, blue item represent data processing steps, and yellow items are expected areas of contribution of this thesis work to the pipeline.

Chapter 2, Microarray Quality Control, addresses specific aim 1, and describes the proposed microarray quality control system, caCORRECT. New algorithms are given for microarray chip artifact detection, artifact aware probe normalization, and artifact aware expression calculation. Chapter 2 gives detailed procedural descriptions of the caCORRECT method, as well as theoretical justifications for the selected approach. In chapter 2, illustrative examples are used to highlight pitfalls of existing methods which are overcome by caCORRECT.

Chapter 3, Quality Control Validation, addresses specific aim 2 by sharing the results of a series of validation experiments that aim to justify inclusion of caCORRECT in to existing workflows. Results show: (1) that previously published biomarker discovery is sometimes correlated or anti-correlated with the presence of chip artifacts, (2) that caCORRECT may be used to increase the reproducibility of biomarker selection during cross-validation, (3) caCORRECT increases the accuracy of existing gene expression calculation methods in the presence of artifacts, and (4) That biomarkers selected from caCORRECT-processed data have a better chance of validation on external samples.

Chapter 4, Development of a Biomarker Based Diagnosis, shares the results of biomarker discovery, aided by caCORRECT, using a case study of Renal Cell Carcinoma (RCC) clinical samples. Biomarkers are first selected from microarray data, and then validated with quantitative RT-PCR and quantum dot immunohistochemistry (QD-IHC). Results suggest that a panel of two markers, NNMT and PRKAB1, can be used in tandem to create an extremely high accuracy (100%, n = 24) RT-PCR based classification system for Clear Cell versus Chromophobe RCC.

Chapter 5, Towards a Quantitative Quantum Dot Methodology, discusses progress made towards the next logical step in this work—a quantitative multiplexed protein-based tissue assay using QDs as reporter molecules. A survey of current issues is given, including QD source separation, characterization of tissue autofluorescence,

differences between QDs, antibody cross-reactivity, and chromatic aberration. For each of these issues, a future outlook is given and the latest work done to address the problem is detailed.

Chapter 6 provides concluding remarks, and highlights concrete deliverables that arose as a result of this dissertation. Finally, an outlook on future work in the field of quality control for biomedical informatics is presented.

CHAPTER 2

MICROARRAY QUALITY CONTROL

The first specific aim of this dissertation is to develop a comprehensive microarray processing algorithm to minimize the effect of common sources of noise on gene expression calculation. Novel contributions include the methods for artifact identification, as well as the concept of artifact-aware normalization and artifact-aware gene expression calculations. Much of the information in this section is an extension or reproduction of the original paper describing caCORRECT which was published in the Annals of Biomedical Engineering [11].

caCORRECT Methodology

The proposed microarray processing algorithm is called caCORRECT, which is short for "chip artifact correction." The general workflow for caCORRECT is shown in Figure 8. caCORRECT operates on probe-level data from any Affymetrix platform microarray and is available for use through a web-interface at cacorrect.bme.gatech.edu. The first step in the caCORRECT algorithm is to normalize each microarray intensity profile to remove global chip intensity biases which may arise due to variation in RNA extraction, amplification, and hybridization procedures. This normalized data is then processed through four rounds of variance calculation, artifact identification, and artifact-aware normalization. After artifacts have been identified, the probe data are fit to caCORRECT's gene expression model, which ignores artifact data. Here, the phrase "gene expression model" is used as a more flexible synonym of the industry term "probe summarization," which belies the two-way relationship between probe data and gene expression. Some readers may be more familiar with this process being called "chip normalization", which is a misnomer that we will avoid altogether.

After fitting data to the gene expression model, users are then given three options: (1) to directly use the gene expression output from caCORRECT's gene expression model, (2) to replace artifact probe data with values imputed using caCORRECT's gene expression model and then send the resulting partially-imputed probe data to an existing third party method of probe summarization, such as RMA, or MAS5.0, or (3) to send caCORRECT's outlier information as a direct input to a third party probe summarization method. This third option is not always available, because it relies on the third party software's ability to take in and use such artifact information appropriately. Each of these steps will be discussed in detail in the following sections.

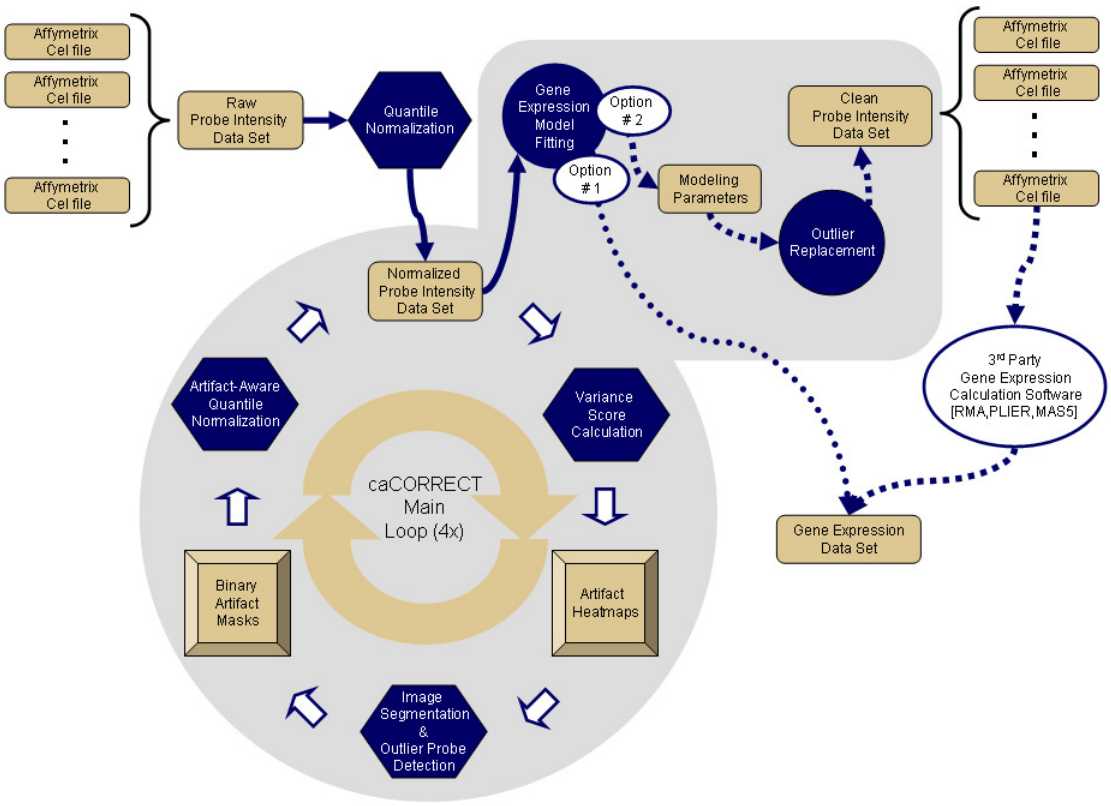

Figure 8: caCORRECT Workflow.
Gold items represent data, while blue items represent data processing steps. The grey region indicates portions of the workflow unique to caCORRECT. These steps are skipped in existing protocols. Raw chip data are the inputs to the system, and both clean chip data as well as gene expression data are the outputs.

Variance Scoring

The cornerstone of caCORRECT's outlier detection is the concept of variance scoring, which produces a value indicating each probe's tendency to be an outlier. Calculation of this score is similar to conducting a t-test to determine whether or not observed probe intensity for a given chip belongs to the observed null distribution of probe intensities for all other chips in the dataset. A key feature of caCORRECT is that this null distribution is updated during each round to include the most up-to-date probe expression data and to ignore any probes on other chips which themselves have been flagged as artifacts. Because of this dynamic updating, it is possible to identify subtle artifacts or pardon false artifacts which may have been misdiagnosed initially.

To estimate this null distribution for each probe on chip j, data from the other chips are analyzed. First, an artifact weighted mean, μ_j, is calculated by using the expression values, x_i, from other chips, ($i \neq j$), in the dataset, and an artifact attenuation factor, α_i. As a default, the attenuation factor equals 1 for spots that have not been identified as artifactual, and 0 for spots that have been identified as artifactual. After this weighted mean is calculated, an artifact weighted deviation score, σ_j, is calculated in a similar manner. These two statistics, μ_j and σ_j, form a description of the null distribution against which x_j is to be compared. The final variance statistic, z_j, for a location is then calculated as the difference between the observed intensity, x_j, and the weighted mean, μ_j, divided by the weighted standard deviation, σ_j.

$$\alpha_i = \begin{Bmatrix} 0: & artifact \\ 1: & not\ artifact \end{Bmatrix} \tag{2.1}$$

$$\mu_j = \frac{\sum_{i \neq j} x_i \cdot \alpha_i}{\sum_{i \neq j} \alpha_i} \tag{2.2}$$

$$\sigma_j = \sqrt{\frac{\sum_{i \neq j} \left[(x_i - \mu_j)^2 \cdot \alpha_i \right] \sum_{i \neq j} \alpha_i}{\left(\sum_{i \neq j} \alpha_i \right)^2 - \sum_{i \neq j} \alpha_i^2}} \tag{2.3}$$

$$z_j = \frac{x_j - \mu_j}{\sigma_j} \tag{2.4}$$

The result of this calculation is a variance statistic, *z*, for each spot on each chip in the study. A high magnitude *z* indicates artifactual *tendency*, while a low magnitude *z* suggests that the spot is to be trusted. Note that a high magnitude score could also be a result of biologically relevant gene expression, but those cases will be generally ignored during the artifact identification procedure described in the next section. A nonlinear scaling procedure is then applied to each *z* to yield *h*, which is a score between 0 and 1 that has been adjusted for the number of chips in the dataset, and is similar, but not equivalent to 1 minus a p-value. This adjustment is made to provide consistent visualization and image processing

$$h = 1 - \exp\left(\frac{-z_j^{\,2}}{n} \right) \tag{2.5}$$

In the above formula, *n* is a scaling factor equal to the ratio of the *p*=0.5 critical point in the student's *t* distribution with degrees of freedom equal to one less than the number of chips, or $\sum_{i \neq j} \alpha_i$, used in the calculation of z_j, to the *p*=0.5 critical point in the Gaussian distribution. Since the number *n* increases with decreasing degrees of freedom, this scaling factor helps ensure that successive rounds of caCORRECT do not progressively and indefinitely shave off more and more outliers until there is nothing left.

This statistic is used as a guideline for identifying regions of chips where artifacts are present as discussed in the next section.

Artifact Segmentation

Once the variance statistic, *h,* has been calculated for each probe on each chip, false-color heat maps of *h*, showing probes in their original spatial layout, are generated to display regions of high noise. For a good quality microarray chip, *h* will represent biological variation in RNA expression for the sample. In this case, *h* will be distributed similarly to white noise throughout the chip. More commonly, however, protocols do not achieve uniform hybridization due to uneven drying, formation of salt streaks, scratching of the microarray surface due to contact with skin or dust, miscalculated hybridization times, or failure to control environmental variables such as ozone [83]. All of these most common mistakes result in visible localized regions of large *h* (artifacts) on the heat map.

Previous versions of caCORRECT included a manual artifact identification tool, but this has been discontinued in favor of an automated batch removal process. Aside from being generally faster and less complicated than manual artifact removal, batch mode has the added advantage of reproducibility. During batch removal, a specialized sliding window method is used to flag probes that meet two conditions: (1) they exist in regions of other high-scoring probes, and (2) they have high scores themselves. These two conditions ensure that most of the obvious artifacts are caught, but that most of the naturally occurring biological variance goes unselected. This automated artifact identification is generally conservative compared to manual artifact identification involving human interpretation of heat maps. Conservative artifact identification is ideal when one considers that caCORRECT is designed to identify spatial artifacts upstream in order to supplement the model-based outlier detection employed by most probe summarization methods. To remove any global chip effects that arise from sample

preparation or amplification, normalization is performed as described in the following section. Figure 9 highlights sample results of artifact segmentation.

Because the intended platform for caCORRECT is a web service, artifact identification has been streamlined for speed and memory efficiency. More computationally intense methods such as active contours, PDE-based methods, or shape matching have been excluded in favor of a quick marching window algorithm that seems to work well for a wide range of data.

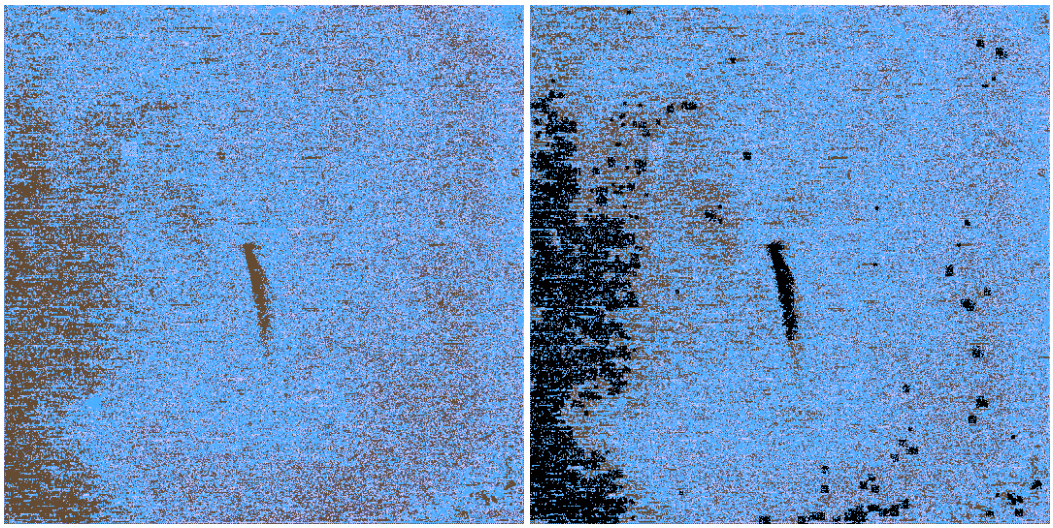

Figure 9: Sample Heat Map and Artifact Segmentation Results.
The left image shows a heat map generated by caCORRECT. The right image shows areas in black where the batch mode of caCORRECT has determined artifacts are present.

Artifact Aware Normalization

Quantile normalization, as described in [84] and recommended by the latest FDA MAQC results[1], reduces bias between microarray samples by forcing the intensity distribution of each chip to be identical. The critical assumption behind quantile normalization is that for genome-wide studies such as those involving microarrays, the number of genes which are invariant to the experimental variables far outnumbers the number of those that are dependent on experimental variables (biomarkers). A set of *n*

distributions is said to be from the same family of distributions (i.e. normal, uniform) if a scatter plot in *n* dimensions of the quantiles from *n* chips results in a straight line. From this it follows that projecting each quantile onto the unit diagonal vector will transform all distributions to one identical distribution. This method is generally appropriate for the microarray problem, where the distributions are poorly defined, and parametric methods break down. The power of quantile normalization comes with a major caveat: if the chips are not from the same distribution, the algorithm will indiscriminately warp the distributions to be the same and proceed as if nothing bad happened. Fortunately, it is a reasonable assumption that high-quality microarray data from a single source on a single platform will follow the same distribution. Unfortunately, this high quality assumption is not valid for much real-world data, where chip artifacts can significantly alter the distribution of intensities on a chip. One bad chip ends up warping the others when quantile normalization is performed, thus compromising the reproducibility of the results. A way to alleviate this problem is to identify artifacts before quantile normalization, and set them aside temporarily. In theory, perfect knowledge of artifacts would allow for perfect correction. This process is called "artifact-aware quantile normalization." In the batch mode of caCORRECT, four iterations of normalization and artifact identification are performed in order to achieve a near steady-state result (data not shown).

To illustrate the invasive effect that artifacts can have on a dataset when quantile normalization is performed, synthetic microarray data were generated in the following manner:

1) Six high-quality chips from the Schultz et al. dataset were chosen, one of which was set aside to receive artifacts.
2) One third of the selected chip was modified by a multiplicative factor of 0.5, representing a low-intensity artifact.
3) A different third of the selected chip was modified by a multiplicative factor of 10, representing a high-intensity artifact.

These six chips were then processed using caCORRECT, and the probe intensities were monitored at the end of each normalization round.

Figure 10 demonstrates the difference between standard quantile normalization and the proposed artifact-aware normalization using artifacts identified with caCORRECT's iterative batch mode on the modified Schuetz et al. dataset [85]. As can be seen in the raw intensities (top panel), the induced artifacts (leftmost and rightmost modes in the blue histogram) caused a differently-shaped distribution than that of the rest of the chips in the dataset. Once the first round of standard quantile normalization was performed (middle panel), the 'warping' of the non-artifacts was clear by the way that each of the distributions were now identically and incorrectly trimodal. After four rounds of artifact-aware normalization were performed (bottom panel), each of the good chips returned to its natural distribution (chips 1 through 4 are obscured behind the pink line, chip 6, since they all have similar distributions). For the (blue) chip with the artifacts, the artifact modes were still clear, but now the remaining non-artifactual data on the chip had been properly normalized.

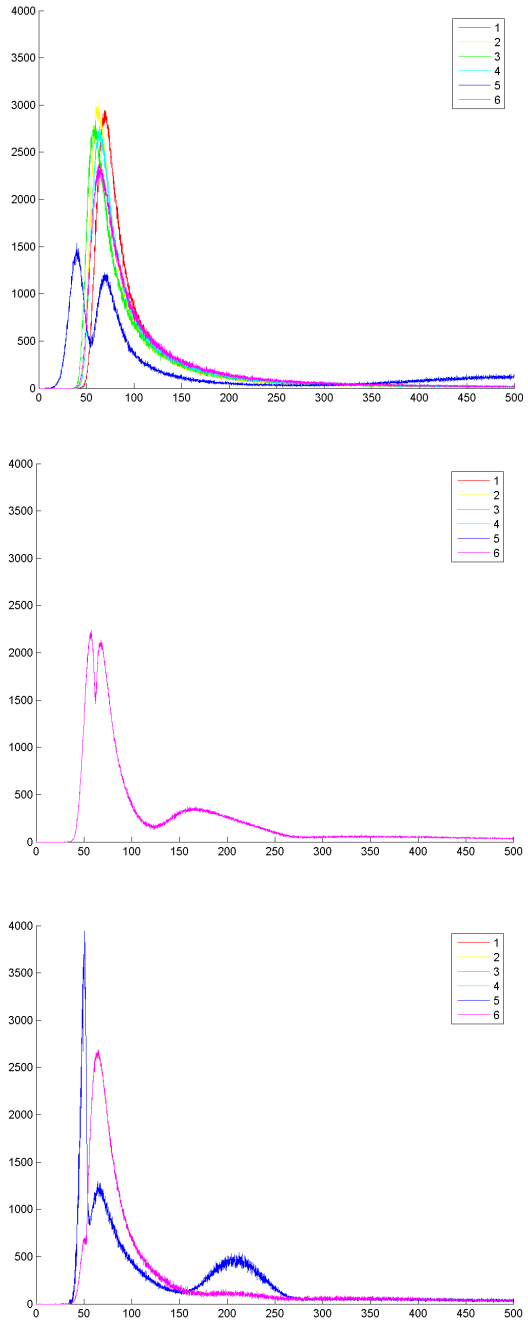

Figure 10: Undesirable Effects of Standard Quantile Normalization and Correction with Artifact-Aware Quantile Normalization Using caCORRECT.
The top image shows the distribution of probe intensities of six microarray chips. The middle image shows the distribution of all chips after quantile normalization. The bottom panel shows the distributions of all six chips after artifact-aware quantile normalization.

Treatment of Identified Artifacts

Integration with Existing Workflow

The general workflow for microarray probe summarization is to input Affymetrix probe-level data and receive gene expression data as an output. This process is analogous to modeling many observations (probe data) of one quantity (gene expression). caCORRECT allows for three different strategies with respect to this general workflow (See Figure 11). The first integration option is to have caCORRECT directly summarize probe data using its own model of gene expression. The second option is to have caCORRECT overwrite data considered to be part of an artifact and then export this 'clean' dataset to a third party method of probe summarization. The third option is to feed caCORRECT's artifact information into existing third party probe summarization methods. Note that the third option may not be possible with all third party software.

To help users interpret results, client software such as RMAExpress or caCORRECT outputs a residual image similar to caCORRECT's variance score heat maps which show the residuals between every probe intensity and the model. Close interpretation of these residuals can reveal how well a particular model fits the data. While the latest version of caCORRECT uses model-based imputation, old versions of caCORRECT simply replaced artifact data for a given probe with the median of that probe's intensity in all other chips of the data set. An example of what such artifact replacement does to the residuals of the RMA model for a real chip is shown in Figure 12. From the figure, it can be seen that median replacement fits the RMA gene expression model fairly well, and that caCORRECT generally does a good job of finding and correcting compact artifacts. Diffuse artifacts, like the one in the upper right hand corner of the chip, are sometimes only partially corrected, due to caCORRECT's conservative approach.

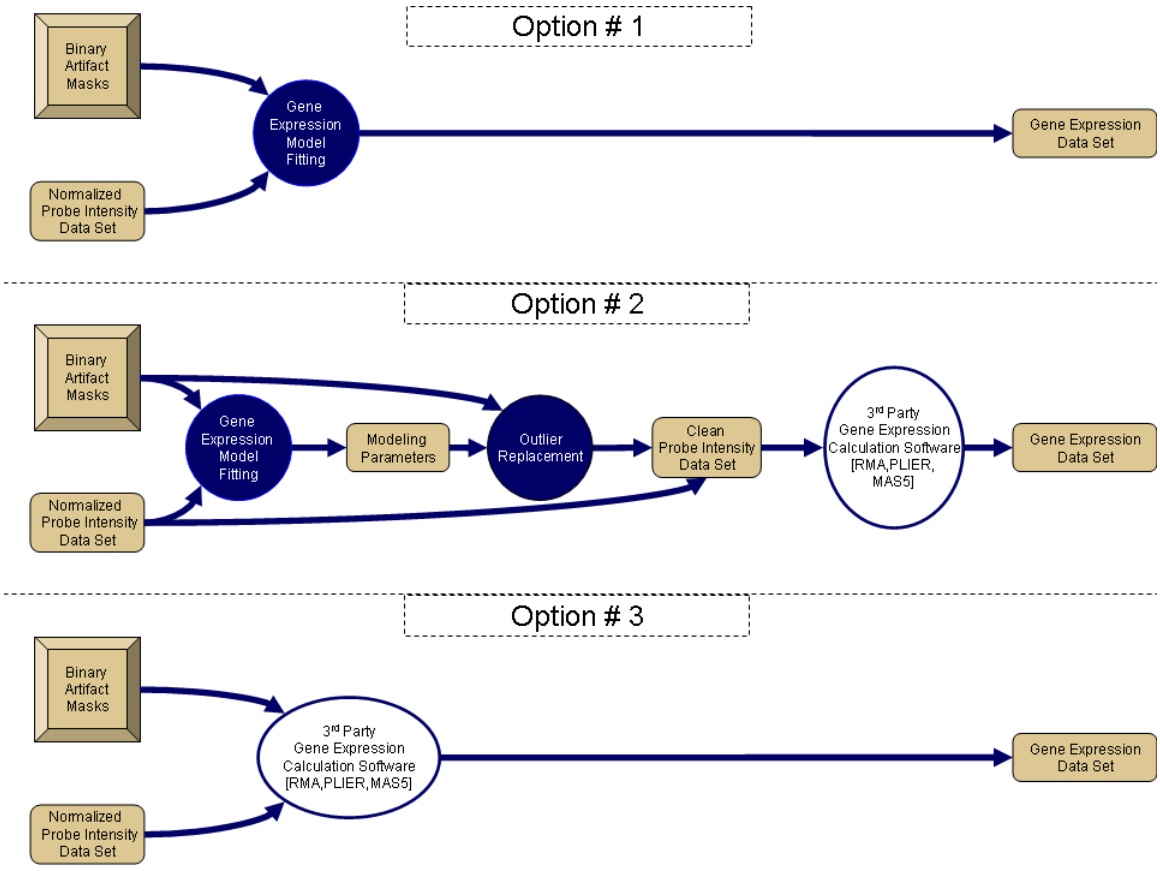

Figure 11: Options for caCORRECT Integration with Third Party Software.
Gold items represent data, while blue items represent data processing steps. White ovals show third party software components.

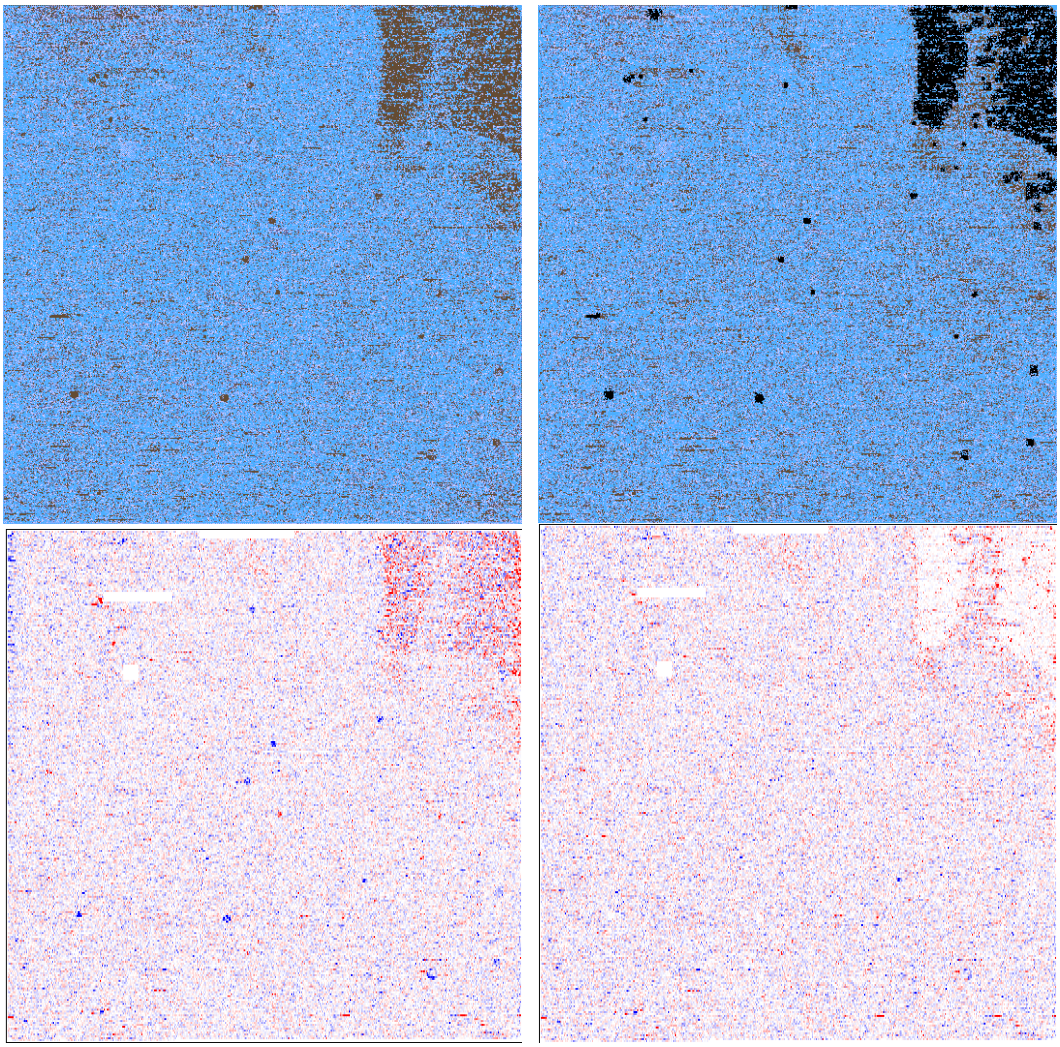

Figure 12: Relationship between caCORRECT and Gene Expression Model Residuals. The top left image shows the heat map generated for an actual microarray chip by caCORRECT, and the results of artifact detection are shown in black in the upper right image. The bottom left image shows the model residuals (positive in red, negative in blue, zero in white) before caCORRECT, and the bottom right shows residuals after artifact data have been replaced with median values (from an older version of caCORRECT). Note that most of the small speckle artifacts are corrected nicely while the more complex artifact is less straightforward.

Modern Gene Expression Model

The current version of caCORRECT uses a model relating probe intensity to gene expression that is mathematically similar to the model used by RMA and others [5, 6, 39]. In this scheme, the artifact-flagged probe-level data is replaced with the best-fit-estimate for that probe given the model and data from non-artifactual probes on all of the chips being processed. To understand how observed probe intensities relate to estimates of gene expression, one must first understand the binding kinetics occurring at the microarray surface.

In a simple model, all copies of a given sequence of the sample cDNA can exist in any one of two states during its application to a microarray: (1) free floating in solution: "unbound" or (2) "bound" to its complimentary strand on the microarray surface (Refer back to Figure 4 for an illustration of this). In an ideal scenario with thin layers of liquid, and abundant binging locations on the chip surface, the transition from the bound to the unbound state can be described with the following kinetic chemical equation.

$$Bound \underset{k_{on}}{\overset{k_{off}}{\rightleftarrows}} Unbound \tag{2.6}$$

In reality, the complimentary binding of two DNA strands is a highly complex process which is not a simple first-order kinetic event, but we will use this simplifying assumption for now, because the downstream conclusion does not rely on this as a fact. After the array has had time to incubate and reach equilibrium, the amount of bound and unbound sample cDNA is described by a ratio of the first order kinetic rates.

$$\frac{Unbound}{Bound} = \frac{k_{off}}{k_{on}} \tag{2.7}$$

There is one notable exception to this formula, which is in the case of an extremely large concentration of sample cDNA. In this case, the target sites on the array can become saturated, and the previous two equations no longer hold. For microarrays,

which measure bound cDNA via fluorescence, the light detector which quantifies the fluorescence will saturate before this point. Detector saturation is easily noticeable, and data which is saturated should be ignored when possible. The "total" cDNA present, which is the quantity ultimately being measured, is the sum total of the bound and unbound cDNA. Using this fact, and some algebraic manipulation, reveals that the amount of cDNA bound to a chip is a fraction of the total cDNA.

$$Total \times \left(\frac{k_{on}}{k_{off} + k_{on}} \right) = Bound \tag{2.8}$$

Because of the binding specificity of DNA, the value of k_{on} will be much larger than k_{off} for complimentary sequences. The opposite will be true for non-matching sequences, and an intermediate balance will be the case for sequences off by only one base pair, such as the so-called mismatch sequences. For our purposes, we assume that (1) the fraction relating total cDNA to bound cDNA exists for every target sequence and it's complementary probe sequence, (2) another, smaller, fraction exists for every target sequence and it's nearly-complementary mismatch probe sequence, and (3) that this fraction is zero for all other probe sequences on the array. Note that Affymetrix arrays are designed to maintain this third assumption and include steps to avoid redundant or overlapping sequences, which may invalidate the assumption.

There are at least two other scaling factors besides this binding fraction which are used to relate the original concentration of RNA in the sample to the observed intensity on the microarray. First, the amplification step, which uses the original RNA from the biological sample to create many fluorescent cDNA copies, introduces a multiplicative gain, g_1. The amount of amplified cDNA is then attenuated by the previously mentioned binding fraction, g_2, to give the amount of cDNA bound. Finally, the process of fluorescence imaging introduces another multiplicative gain, g_3, relating the amount of

bound cDNA to the fluorescence intensity. These three factors can be lumped together to give an overall probe affinity factor, *a*.

$$[RNA] * g_1 * g_2 * g_3 \approx fluorescence \tag{2.9}$$

$$g_1 * g_2 * g_3 = a \tag{2.10}$$

In caCORRECT, observed microarray intensity values are modeled as a multiplicative combination of target RNA abundance (gene expression) and probe-specific effect (probe affinity). The model is given as:

$$x_{b,p,j} = \theta_{p,j} a_{b,p} + \varepsilon_{b,p,j} \tag{2.11}$$

where $x_{b,p,j}$ is the observed intensity for the b^{th} probe in the p^{th} probe set on the j^{th} chip, $\theta_{p,j}$ is the gene expression term corresponding to initial RNA concentration, $a_{b,p}$ is the lumped probe affinity term, and $\varepsilon_{b,p,j}$ is the additive error term which accounts for both instrument noise as well as nonspecific binding. For a detailed discussion of the noise model, and an explanation of the solution to this gene expression model, refer to Appendix A. The model is both regressive and generative in the sense that it may be used to estimate the parameters $\theta_{p,j}$ and $a_{b,p}$ from a given dataset, and it may also be used to impute synthetic intensity data $\hat{x}_{b,p,j}$ given a set of $\theta_{p,j}$ and $a_{b,p}$. As the last step of caCORRECT, all probe data determined to be part of an artifact is replaced by this imputed intensity value, $\hat{x}_{b,p,j}$.

Applying in the Gene Expression Model

The Affymetrix microarray is redundant, in that there are usually 40+ probes on a single array (20+ PM, 20+ MM), that target a single gene. Solving the model equations is still underspecified for a single chip; i.e. there are 40 observed intensities (40 equations), but 40 probe affinities and 1 gene expression term to estimate (41 unknowns). Adding a second chip to the batch doubles the number of equations but only adds one new

unknown (a gene expression to estimate) since probe affinities are a sequence and protocol specific value that should be conserved across chips in a single dataset. In the case of more than one chip, the model becomes increasingly over specified, and it can be solved using standard least-squared-error techniques (See appendix A). This over specification allows caCORRECT and other models the option to throw out large amounts of data and still reliably estimate gene expression. The benefit that caCORRECT provides is improved knowledge of which data to discard.

Figure 13 represents a case study using residual images to compare how caCORRECT and RMA react to the presence of a scratch on a chip from the West dataset [86] (also shown in Figure 9). The layout of the Hu-6800 chip is such that all probes which make up a single probe set are arranged in contiguous regions on the microarray (See "retired format" in Figure 3). This property allows for easier visual interpretation of residual images, and is why this dataset was chosen for this case study. The leftmost panel of Figure 13 shows the residuals produced by the caCORRECT model without any attempt to identify artifacts or ignore outliers. The blue (negative residual) regions which surround the main red (positive residual) scratch demonstrate the ambiguity which arises when outliers are present. The dilemma for the modeling algorithm is that it doesn't know if the red data are artificially high, or if the blue data are artificially low. As a result, the optimal uninformed model splits the difference, and assumes that all data points are slightly wrong. In contrast to the uninformed model, this disambiguation is easily made by caCORRECT, which identifies the actual scratched region as bad, and the surrounding data as good by *h* score alone (see the heat map of *h* in Figure 9). caCORRECT's segmentation algorithm further disambiguates the scratch from the background. The residuals which are calculated after caCORRECT has identified artifacts are shown in the middle panel of Figure 13. At this step, it is clear by the white region surrounding the scratch that caCORRECT has successfully disambiguated the decision, deciding that the high-intensity data of the scratch are to be ignored, and the

resulting model uses only the remaining probe data to estimate gene expression. The rightmost panel of Figure 13 shows the final result produced by RMAExpress without any intervention by caCORRECT. RMA includes its own outlier detection, but RMA's detection is not informed by spatial location, and thus cannot recognize a scratch as such. The figure suggests that RMA does a fair job of disambiguating the scratch from the surrounding data, but some areas of blue still exist, suggesting that the gene expression values for these probe sets have been overestimated. These problems have been almost completely avoided by caCORRECT.

Another perspective on this same case study is shown in Figure 14, which shows the result of model fitting before and after caCORRECT quality control to identify artifacts. The data from four probes are displayed for each of the 49 chips in this dataset. In both cases, the chip with the highest gene expression is the one discussed earlier and shown in Figure 13. Regression lines for each probe are also provided for reference. Points appearing above their line give a positive residual, while those below are considered to have a negative residual. Without proper knowledge that the cyan-colored probe on the scratched chip is part of an artifact, the best-fit model estimates log gene expression for the questionable chip near 8.3. With caCORRECT-generated knowledge, the log gene expression is estimated nearer to 7.4, a change of an entire order of magnitude. The proximity of the remaining data to their respective regression lines is another indication that caCORRECT chose correctly.

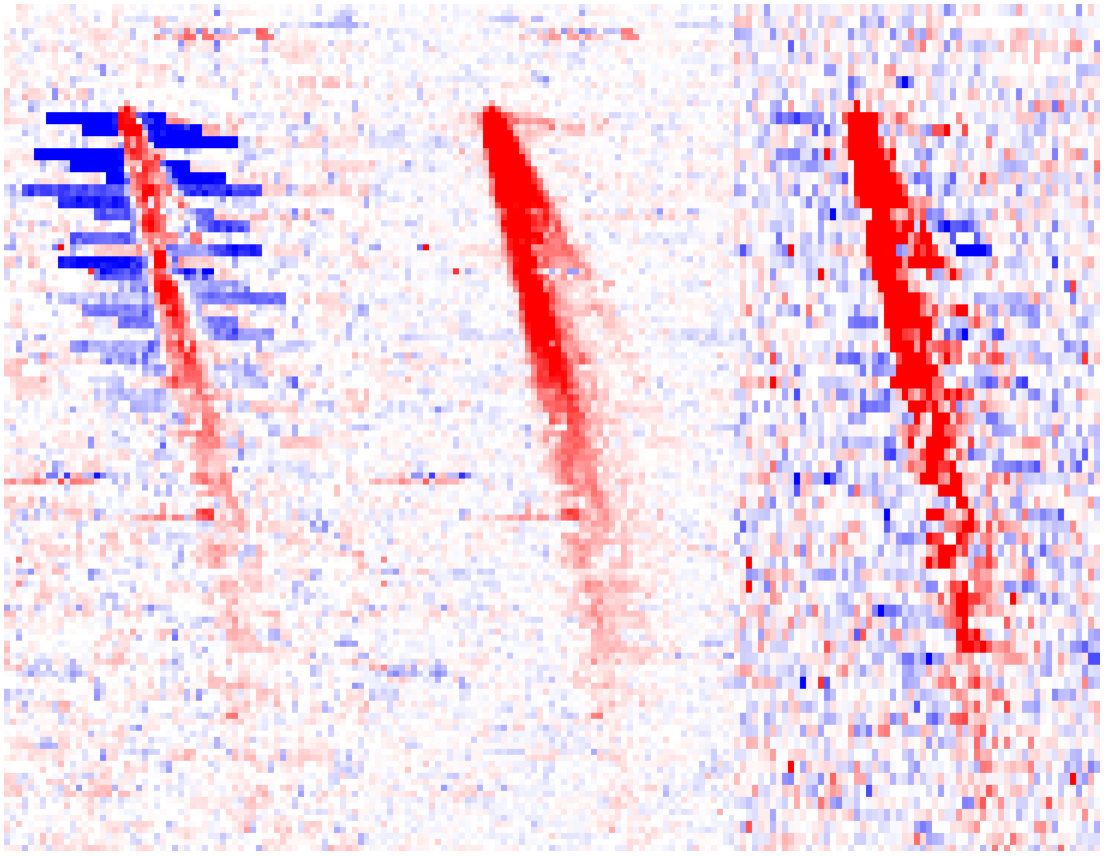

Figure 13: Effect of Scratch Artifact and Removal on Residual Images.
Red color indicates positive residuals, blue color indicates negative residuals, and white indicates good fit to the model. Blue color surrounding the red scratch reveals poor model fit to the data which is likely to cause an overestimate of gene expression for these probes. (left) Initial residual produced by caCORRECT. (middle) Residual produced by caCORRECT after artifact flagging. (right) Residual produced by RMA. The figure indicates that caCORRECT is more robust to this scratch than RMA is. The reduced resolution for the right panel is explained by the way that RMA handles pairs of perfect match and mismatch probes together when calculating residuals.

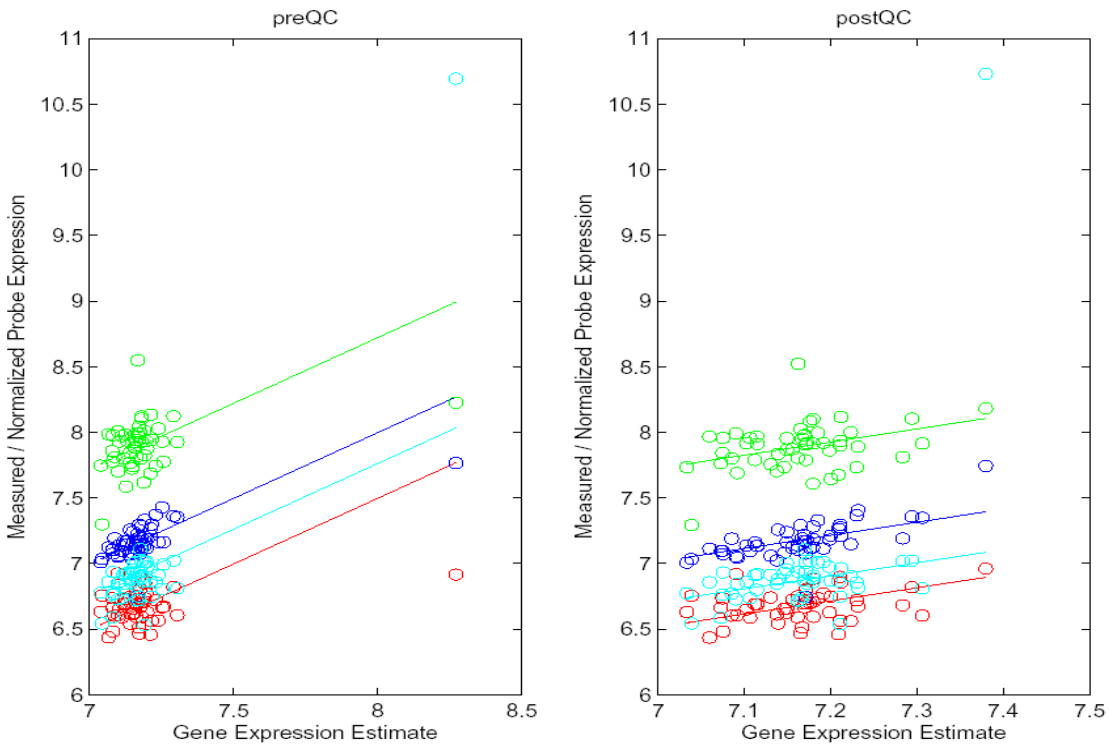

Figure 14: Effect of Scratch Artifact and Removal on Gene Expression Estimate. Four probe sequences for this probe set are shown, each with its own color. Lines represent the best-fit line relating probe expression to gene expression. Data from the chip with the scratch are the rightmost data points

Summary

In this chapter, we have described the methodology of caCORRECT, and discussed many theoretical and practical shortcomings of existing methods of outlier detection and global chip normalization.

Using a small 6-chip dataset as an illustrative case study, we have shown how the existing gold-standard normalization method, quantile normalization, can have a severe negative impact on the data quality it attempts to improve. Furthermore, we provide a method whereby this warping phenomenon can be almost completely avoided, without compromising the original intent of the quantile normalization algorithm.

We also demonstrate, using the example of a prominent scratch on a real clinical microarray, the superior artifact finding ability of caCORRECT as compared to existing model-based methods such as RMA. While existing technologies, including robust modern chip layouts and robust statistical outlier detection algorithms, do perform satisfactorily enough for microarrays to be considered a clinically viable technology, caCORRECT represents a significant technical improvement in microarray data analysis. The next chapter discusses further empirical validations of caCORRECT.

CHAPTER 3

QUALITY CONTROL VALIDATION

The second specific aim of this dissertation is to investigate the effect of microarray noise and quality control on the reproducibility of gene biomarker discovery. Direct comparisons of caCORRECT to existing methods are difficult because, with each of the methods mentioned earlier, noise removal and gene expression calculation are either inseparable or not integrated. This therefore limits experimentation to the form of comparing A) using caCORRECT and B) not using caCORRECT before proceeding to gene expression calculation by the most widely accepted methods, RMA and MAS5.0. In this chapter, a series of experiments are conducted to showcase the various ways in which caCORRECT can improve microarray results. Some of the results in this section are reproductions of my contributions to publications in the Annals of Biomedical Engineering [11] and Life Science Systems and Applications Workshop [12].

Overlap of Presumed Biomarkers and Chip Artifacts

This first investigation was actually more of a justification than a validation of the use of caCORRECT. Before attempting to improve any results with caCORRECT, we first made sure that the problem that caCORRECT addresses (chip artifacts) was linked to the problem that we wish to address with caCORRECT (successful translation and reproducibility of results). Accordingly, in this section, we answer the question: “Is there a relationship between published microarray results in literature and chip artifacts?” To answer this question, we conducted a survey of existing high-impact publications which had also provided access to probe intensity data. In these publications, we investigated whether their proposed biomarkers could be correlated with to chip artifacts. Although correlation does not imply causation, it at least allows the possibility that solving the problem of microarray artifacts may have an affect on biomarker selection.

Methods

To measure if there is a relationship between published biomarkers (identified before the existence of caCORRECT) and artifacts flagged by caCORRECT, overlap was observed between these two events using a combination of literature mining and post-hoc analysis. Previously identified biomarkers from third-party analysis were manually mined from the literature and mapped back to their spatial locations on the chip. These locations were then overlaid with the location of artifacts discovered during the QC process using the binary masks produced by caCORRECT. To produce a *whole dataset* mask, showing the number of artifacts found at each probe throughout the entire dataset, the *N* artifacts masks from each of the *N* chips were overlaid (summed together). The final result describes the locations and number of artifacts as well as the locations of biomarkers.

In order to analyze this data quantitatively, a comparison was made between the number of artifacts found on biomarker probes, and the number of artifacts found on all other probes on the chip. First, two discrete probability density functions were estimated: one for the set of biomarker probes, f_b, and another for the rest of the probes on the chip, f_r. These density functions were estimated directly from the observed overlap frequencies and represent the set of probabilities ($f_b(i)$ or $f_r(i)$) of encountering i artifacts at a random probe. After the two probability density functions were estimated, they were then compared using the dissimilarity measure proposed by Vajda given below[87]:

$$\sum_{i=0}^{N} \left| f_b(i) - f_r(i) \right| \tag{3.1}$$

The distribution of this dissimilarity statistic is not well defined, and varies based on the number of biomarkers as well as the artifact coverage of the dataset. In order to estimate the significance of any observed dissimilarity, an ad-hoc null distribution of dissimilarity was generated for each new dataset by permuting the class labels for the biomarker status of each probe many times. The distribution of the dissimilarity measure

across the permutations serves as a basis for estimating the significance of the dataset's true dissimilarity value. Highly significant dissimilarities are an indication that the association between biomarkers and artifacts for a dataset is nonrandom.

Results

Visualization of chip regions with high numbers of artifacts can reveal patterns of possible chip mishandling. Figure 15 shows a visualization of all of the artifacts found by caCORRECT in the dataset published by Shipp et al. overlaid with the locations of biomarkers claimed by the original authors [88]. The first observation to be made from this visualization is the remarkable preference of artifacts to occur on the right hand side of the array. Without detailed observations of chip operators, it would be difficult to draw strong conclusions as to the origins of these errors, but it is easy to image a scenario where a novice operator has a tendency to touch the array surface with their right thumb in such a way as to consistently degrade the quality of that side of the array.

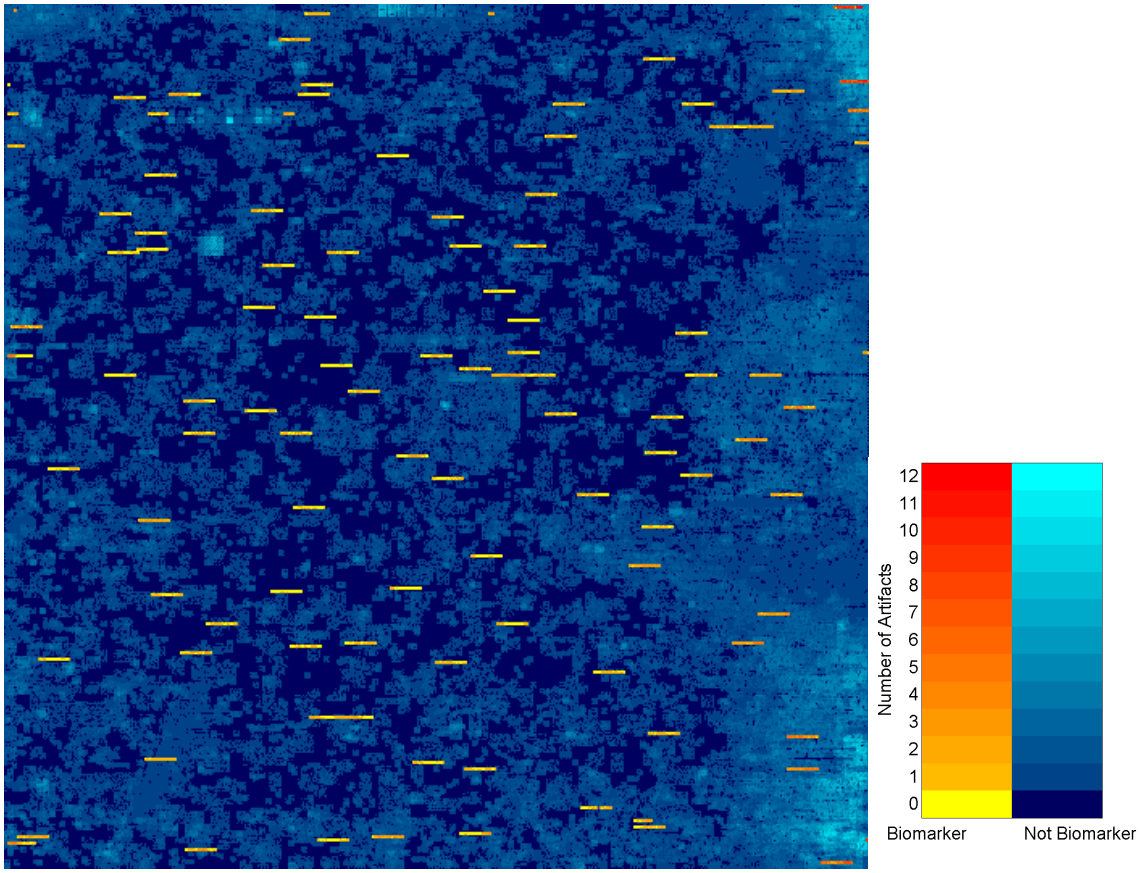

Figure 15: Overlap of Active Artifact Regions with Biomarkers.
Artifacts seem to be concentrated on the right hand side of the chip. Biomarkers can be found preferentially in these most artifact-rich regions of the chips.

Visualizing and quantifying the overlap of biomarkers and artifacts is a critical step in forming conclusions about the quality of microarray data and the reliability of biomarker selection. However, interpreting the results of these analyses can be somewhat controversial. Figure 16 shows analysis results derived from three manuscripts and their corresponding datasets, which exhibit various degrees of overlap between biomarkers and artifacts. Three scenarios are evident.

In the first scenario, as seen in the Schuetz et al. dataset [85], genes of interest were uniformly and independently distributed with respect to outlier probes. This suggests that outlier data were unable to bias the gene selection process. However, upon further review, one might question the validity of these supposed biomarkers which were identified based on noisy data, in terms of both false positives as well as false negatives.

In the second scenario, as seen with the West et al. dataset [86], the genes of interest were located preferentially in the less noisy portions of the microarray. One might conclude that these biomarkers are 'better' because they have been derived from less noisy data. However, many 'good' biomarkers may have been excluded simply because they are hiding in noisy regions of the chip.

The third scenario, from Shipp et al. [88], when biomarkers co-occurred with outlier probes, is not conclusively bad. It could be that the detection of outlier probes was biased by the true differential expression of a biomarker and not the reverse. Proper interpretation of these overlap results requires careful examination of the exact methods of quality control, outlier detection, and biomarker detection and how they may affect each other. Such thorough analysis is impractical for even the most carefully written manuscripts, because of generally irreproducible methodologies. Due to this limitation, we focus our further investigations on our own controlled experiments.

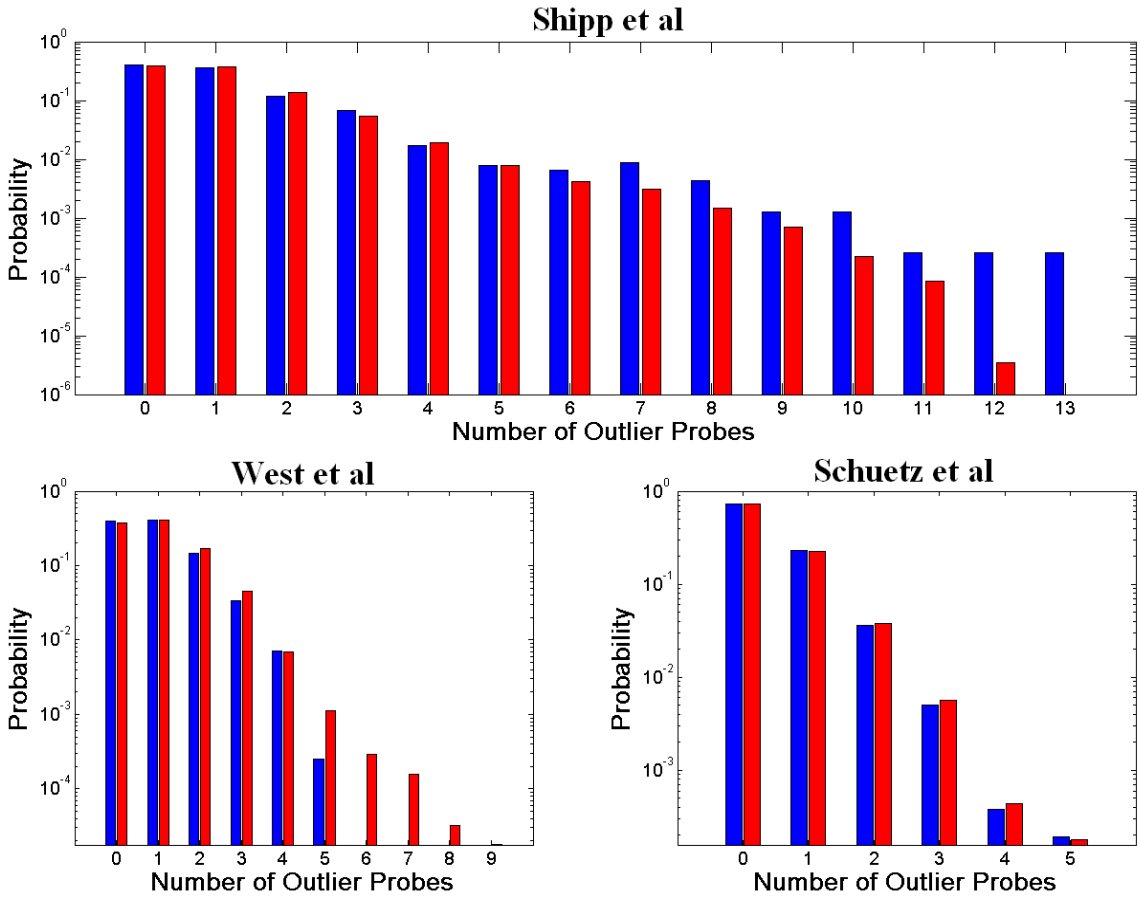

Figure 16: Selected Empirical Probability Density Functions.
Blue represents genes of interest and red represents genes not of interest. (top) The Shipp et al. Diffuse B-cell Lymphoma dataset [88] showed a significant ($p < 0.001$) trend of co-localization. (bottom right) The Schuetz et al. renal dataset [85] showed no significant overlap ($p > 0.05$). (bottom left) The West et al. Breast Cancer dataset [86] showed a significant ($p < 0.001$) trend against co-localization. This figure has been modified from its original version, presented at the Life Science Systems and Applications Workshop[12].

caCORRECT and Reproducibility of Feature Ranking

A list of biomarkers is the common result of a microarray experiment, and thus is the common target of many studies attempting or commenting on reproduction of microarray experimentation [22, 25, 26, 29, 89, 90]. A ranked list is a special case of biomarker selection that lends itself to more precise comparison than unranked lists. This section investigates how caCORRECT influenced the repeatability of ranked lists generated from two halves of a large microarray dataset.

Methods

To determine the ranked gene list for each dataset, one-dimensional linear Support Vector Machines (SVM) was used. SVM was chosen because of its ability to generalize well in the face of noise, and its stable error estimation, which make it ideal for selecting ranked lists of candidate biomarkers. The SVM classifier uses a Gaussian bolstering kernel with a bolstering radius of 1.4826 and 100,000 points of Monte Carlo integration for error estimation. More information on the implementation of SVM, its application to microarray study, and sensitivity to quality control, can be found here [8, 9, 64, 91, 92]. The classification error output by the SVM was used as a basis to rank each gene.

To measure caCORRECT's effect on the reproducibility of microarray experimental conclusions, ranked lists of biomarkers were compared from two independent datasets obtained by splitting one large dataset in to two halves. This split ensured isolation of confounding variables such as lab technique or array platform. The goal was to find little change between ranked gene lists of two different datasets—indicating reproducibility of findings. To create these independent datasets, the 49 chip dataset published by Huang et al. [93] was randomly split into two non-overlapping data subsets. In this way, one large 24v25 dataset was split into subsets of 12v12 and 12v13.

Each of these subsets was then used as input for a support vector machines gene ranking routine twice; once without any QC, and once after experiencing artifact flagging and data replacement from caCORRECT. RMAExpress was used for all gene expression calculations in this study. Figure 17 shows a schematic for this experimental design.

The above process produced four different lists of ranked biomarkers. These four lists were then compared to each other in pairs to determine reproducibility of findings. To compare lists, first, the ranks were transformed to the log2 scale to reflect the decrease in marginal discriminatory power with an increase in rank. Any gene which occurred in the top 100 of either list was chosen for further analysis. For each of these chosen genes, the difference in log2 rank for that gene between the two lists was calculated. The distribution of this difference statistic thus describes the similarity of the two ranked lists. The entire data-splitting, feature-ranking, and rank-comparing procedure was then repeated many times to get a better estimate of the true distribution of the difference statistic.

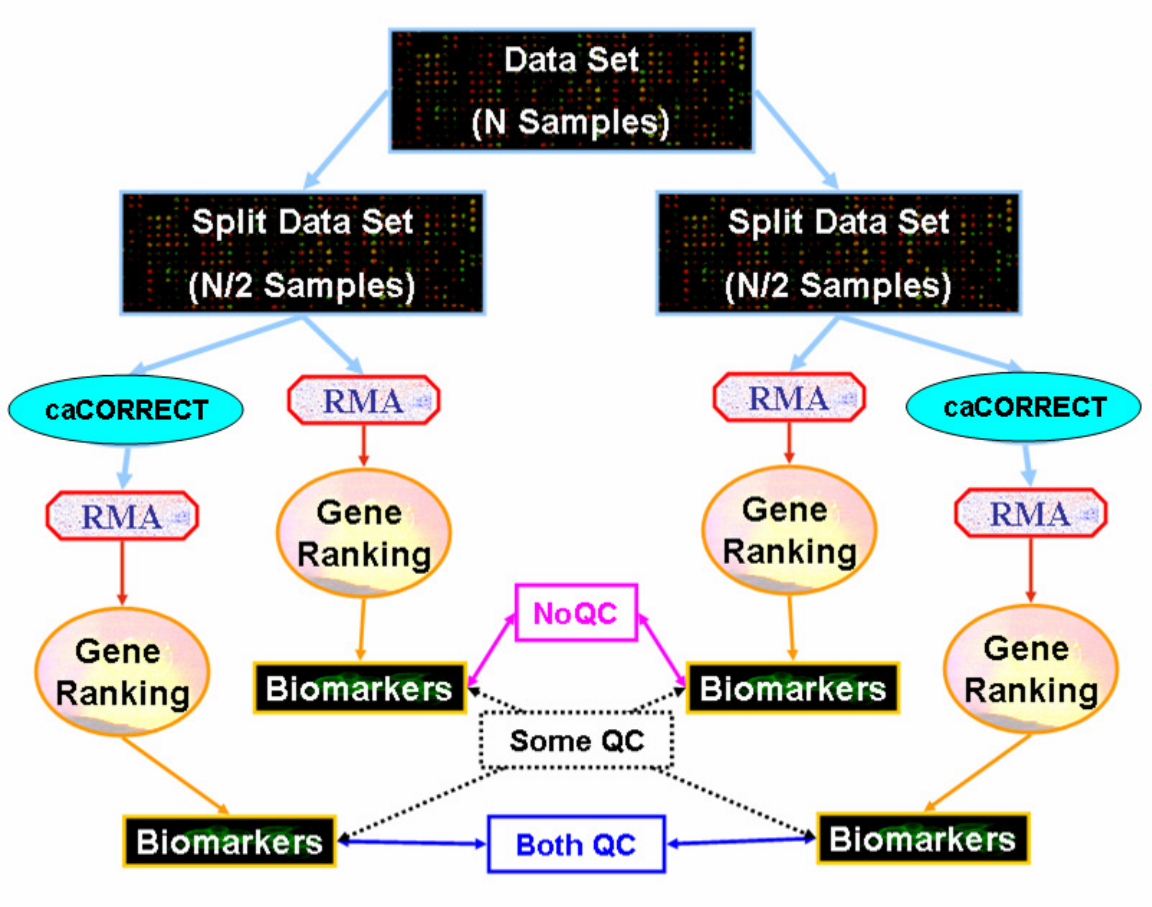

Figure 17: Illustration of Workflow for Biomarker List Comparison.
Datasets consisting of probe-level microarray data are shown in light blue, expression level microarray data is in red, and ranked lists in orange. The three comparisons made here correspond to the results shown later in Figure 18. This figure reproduced from the 2007 ABME paper [11].

Results

Results from the independent splitting of the Huang dataset are summarized in Figure 18. It can be seen that using caCORRECT on both halves of the dataset improved the reproducibility of ranked gene lists in two ways. First, distribution plots show an increase in the number of genes whose rankings did not change more than two-fold; i.e. more gene ranks were conserved. Secondly, the mean of rank differences after caCORRECT was significantly lower than that obtained without caCORRECT ($p<0.05$).

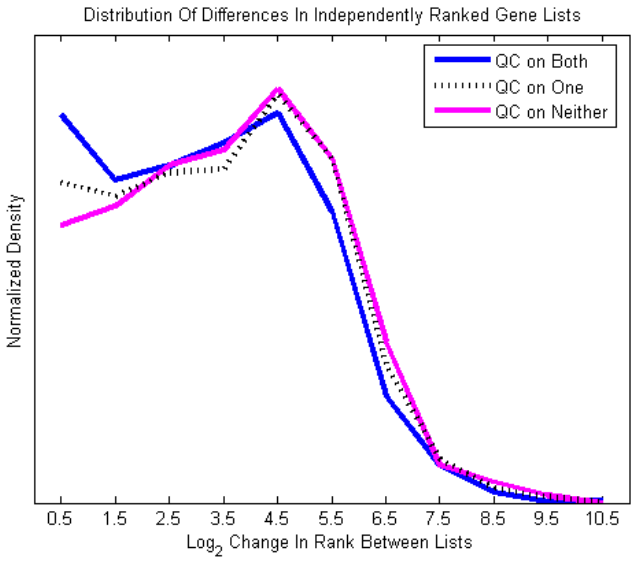

Figure 18: Effect of caCORRECT on Similarity of Ranked Gene Lists during Cross Validation.
This figure reproduced from the 2007 ABME paper [11].

Effect of Applied Artifacts and Preprocessing on Gene Expression

With the motivation that there is (sometimes) a link between artifacts and biomarkers, and that caCORRECT preprocessing can increase the reproducibility of feature lists, we now remove biomarker selection and classification from the pipeline, and instead investigate how caCORRECT improves the accuracy of the microarray itself as a clinical profiling tool. More precisely, this next investigation quantifies how

caCORRECT improves the accuracy of gene expression, which is the driving input to both biomarker selection and clinical decision making.

Determining the accuracy of a microarray in the traditional sense usually involves comparison to other arrays of or PCR measurements of the same biological samples[1] or measurement of known "spike-in" transcripts[5, 46, 59, 94]. Instead of these costly methods, we instead chose to compare microarrays to themselves by simulating new versions of arrays with applied artifacts. Only high quality arrays were chosen so that each array may serve as its own gold-standard.

Methods

In order to quantify the ability of caCORRECT to improve the accuracy of microarray gene expression, and thus the reproducibility of array dats, a set of otherwise high-quality breast cancer microarray data (selected for quality by the MAQC-II Consortium) were altered with a series of randomized synthetic artifacts and then processed with caCORRECT. Artifacts were originally generated for the MAQC-II Jones et al. study on classifier performance in the face of artifacts, and were created blindly with respect to caCORRECT and its developers.

Two types of artifacts from Jones et al. were investigated here: (1) the h90 "black hole" artifact in which an elliptical region of the microarray had probe intensities lowered severely, and (2) the s90 "hot spot" artifact in which an elliptical region of the microarray had probe intensities raised severely. Such elliptical artifacts represent one of the two most frequently observed artifacts in microarrays. The other type of artifact is more diffuse in nature, and tends to appear near the edges or corners of arrays, most likely due to fluidics and drying issues. For the Jones et al. study, we were able to use caCORRECT to identify and transplant this second type of artifact from actual chips onto the chips in the Jones study. We did not, however, use these caCORRECT-generated artifacts in the

analysis of caCORRECT's ability to remove artifacts for the purpose of scientific integrity.

For each of the Jones-created elliptical artifacts, the orientation and location of the artifact region was altered randomly for ten times on each of the 49 chips of the validation set designated by the original microarray data suppliers, Hess et al. [95]. For each of the 10 altered chips, gene expressions were calculated both before and after caCORRECT's complete artifact detection and value imputation were applied. Each of these estimated gene expressions was then compared to the "true" gene expression values obtained from the respective original, unaltered chip to yield an error value representing the deleterious effect of the artifact on gene expression estimation for each probe set. The errors for each probe set (~8000), each chip (49), and each artifact replicate (10) were then pooled together to form two distributions of the error function for each of the two artifact types: one for unprocessed data, and one for data cleaned with caCORRECT. For each chip, gene expression data for all probes were determined using MAS5 in Expression Console and the R implementation of RMA.

Results

The effect of applied quality insults on gene expression was monitored for two of the most popular probe summarization methods: RMA and MAS5.0. The agreement of gene expressions before and after induced artifacts is shown in Figure 19. Two phenomena are immediately observable:

First, for the "black hole" artifacts which lower probe intensities on the microarray, the MAS5.0 algorithm had the tendency to call many of the genes 'absent', and report the gene expression abnormally low. caCORRECT was able to almost completely reverse this trend, and was generally able to rescue the true gene expression.

Second, for the "hot spot" artifacts which raise probe intensities on the microarray, the RMA algorithm had the tendency to underestimate gene expression and

lose accuracy for the genes most highly expressed in the sample. This was most likely a result of the issues related to quantile normalization discussed in Chapter 2. This phenomenon also happened to RMA to a lesser extent for the black hole artifacts. The robustness to black holes was most likely due to a natural dilution effect caused by the relatively large number of low-expression (in many cases, mismatch) probe sets on a microarray. Chips processed first with caCORRECT and then with RMA did not seem to exhibit either of these warping behaviors.

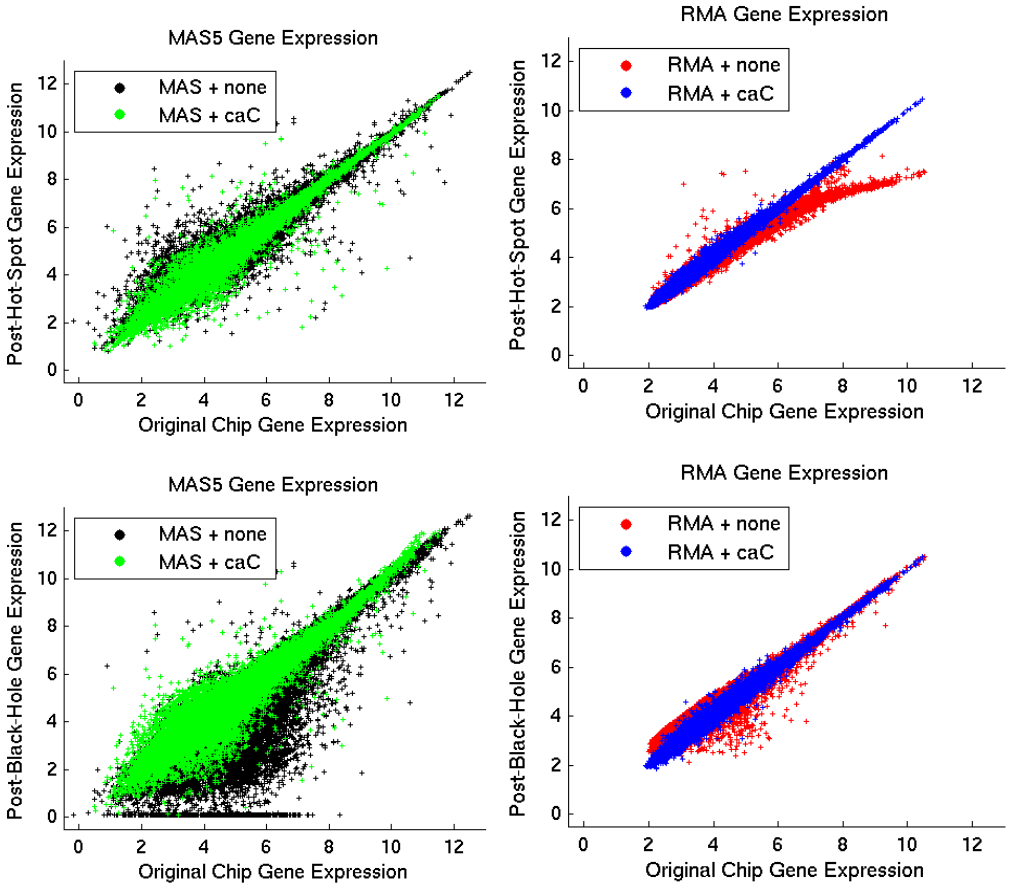

Figure 19: Scatter plots of Gene Expression after Quality Insult Versus Original Gene Expression.
Data shown are for one representative chip, and all probe sets on the HG-U133A platform. Gene Expression is calculated either independently with MAS5 or with RMA as part of a batch containing the 81 chips in the training set. caCORRECT normalization is performed as part of a batch with the 81 chips of the training set. Units of gene expression are on the scale of the natural log of probe intensity. caCORRECT improves gene estimation in all cases, as exhibited by its scatter plots being closer to a line with unitary slope.

For each combination of preprocessing and probe summarization, the root mean squared error (RMSE) of the gene expression estimate was calculated by assuming the ground truth to be the gene expression obtained from the original unaltered chips. Results confirmed the scatter plot evidence, and show that MAS5 performed poorly on the subtractive "black hole" artifacts, while RMA performed poorly on the additive "hot spot" artifacts. In both cases, preprocessing with caCORRECT was able to reduce the effect of the artifacts as seen in a reduction of RMSE. Figure 20 and Table 3 summarize these findings in terms of the distribution of RMSE across all probe sets on the chip.

Table 3: Effect of Artifact Type and Preprocessing Procedure on Precision and Reproducibility of Gene Expression.
Shown are the mean values for the curves shown in Figure 20.

Artifact type	**Preprocessing**	**Gene Expression Method**	**Mean RMSE**
hot spot	none	MAS5.0	0.31
hot spot	caCORRECT	MAS5.0	0.25
hot spot	none	RMA	0.32
hot spot	caCORRECT	RMA	0.11
black hole	none	MAS5.0	0.96
black hole	caCORRECT	MAS5.0	0.35
black hole	none	RMA	0.38
black hole	caCORRECT	RMA	0.15

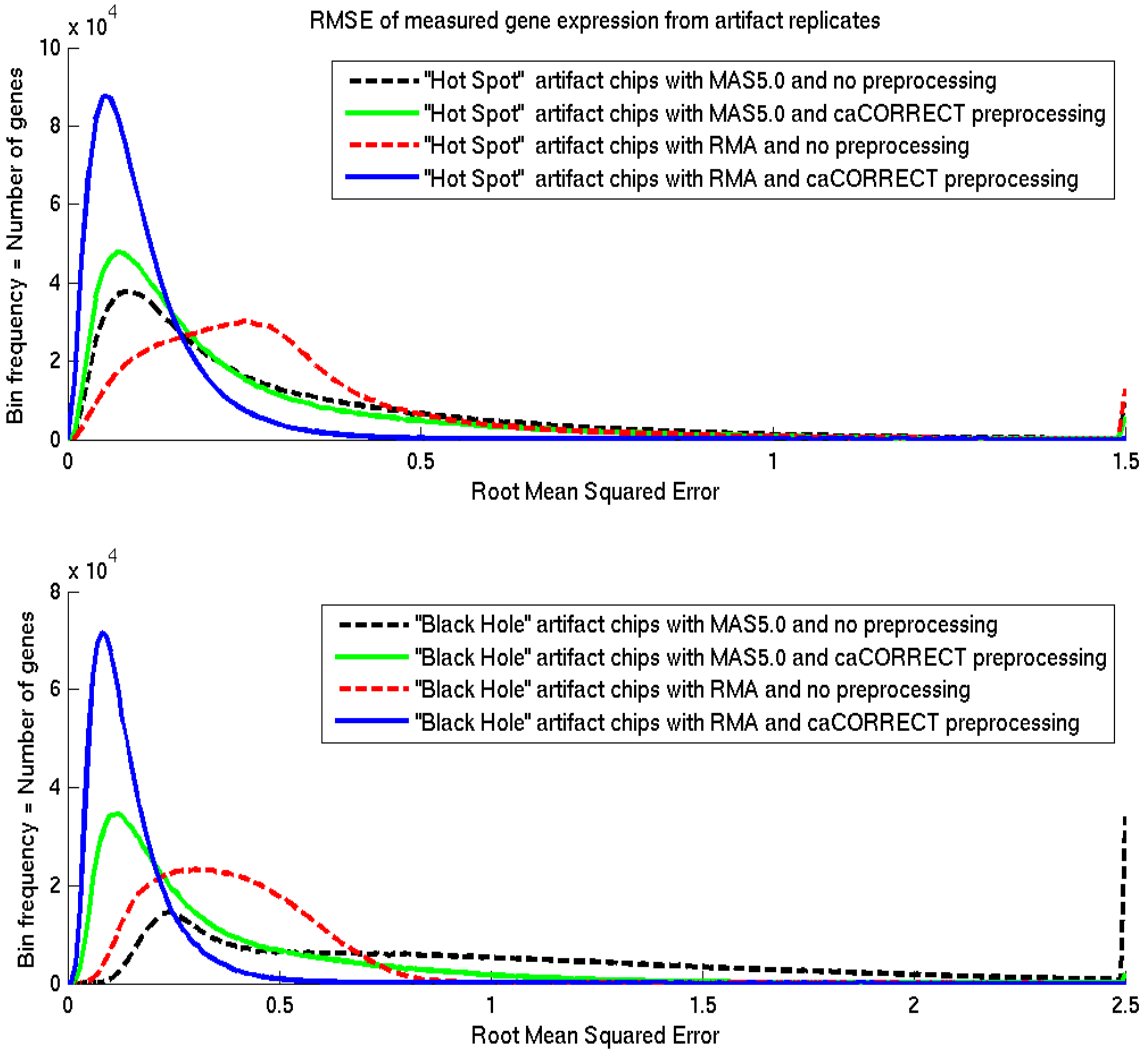

Figure 20: Effect of Artifact Type and Preprocessing Procedure on Error of Gene Expression Estimation.
Data shown are for all 49 chips in the validation dataset, and all probe sets on the HG-U133A platform (composing the histogram bin counts). Root Mean Squared Error is calculated for each probe set separately across the 10 Monte Carlo artifacts. RMSE is calculated assuming the ground truth to be the gene expression values derived from the original, unaltered chips using each respective combination of preprocessing and summarization method. Units of gene expression are on the scale of the natural log of probe intensity. caCORRECT improves the reproducibility of these chips by reducing RMSE as evidenced in the left shift of the histograms.

Clinical Validation Pilot Study

The final and most clinically relevant validation of caCORRECT is testing to see if caCORRECT has an appreciable affect on the reliability and reproducibility of the translational bioinformatics pipeline. This validation began with a PCR pilot study of microarray-derived biomarkers that changed status with quality control.

Methods

Three separate ranked lists of biomarkers were produced from the RCC Microarray data [85] according to Figure 21: (1) before any quality control, (2) after processing with caCORRECT, and (3) after removing two chips deemed to be unacceptable after processing with caCORRECT. The resulting ranked lists are referred to as pre-QC, post-QC (all), and post-QC (trim) lists respectively. All gene expression calculations were obtained using the RMA algorithm, using RMAExpress. When using RMAExpress, all data were processed using both the quantile normalization and background correction features included in the software. Ranked gene lists for each dataset we created with one-dimensional linear Support Vector Machines (SVM). The SVM classifier uses a Gaussian bolstering kernel with a bolstering radius of 1.4826 and 100,000 points of Monte Carlo integration for error estimation. The classification error output by the SVM is then used as a basis to rank each gene.

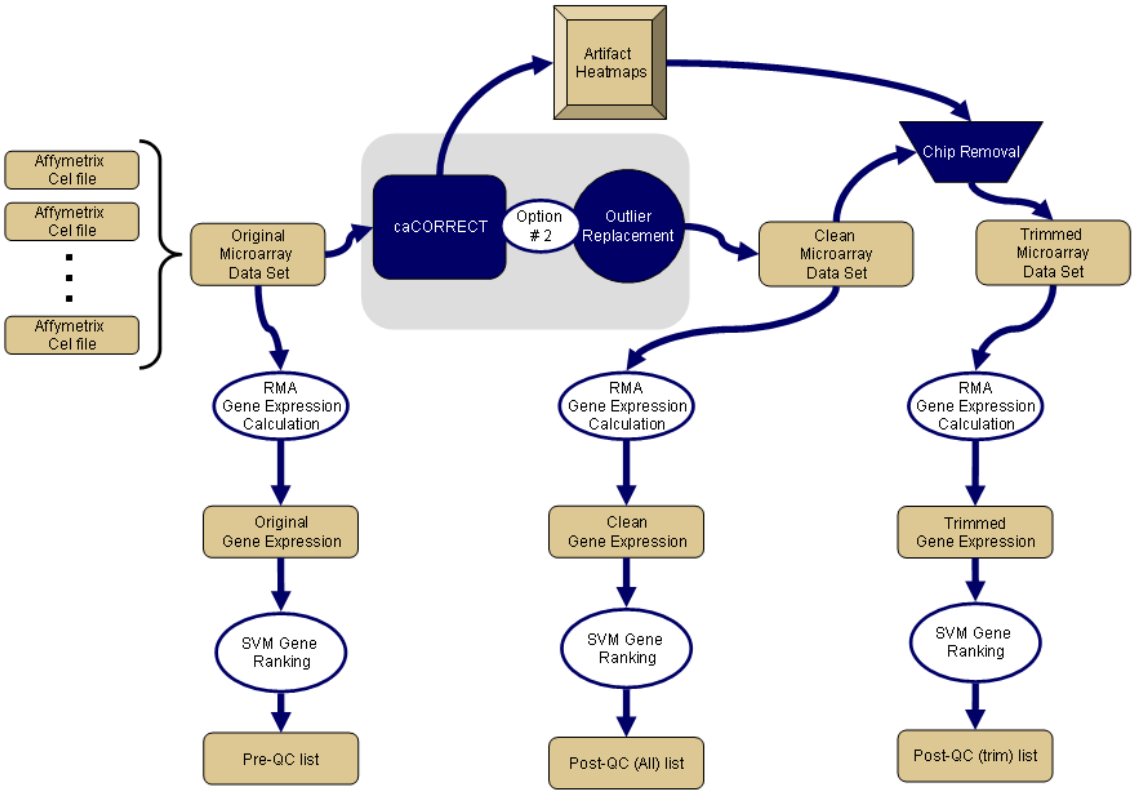

Figure 21: Workflow for Clinical Pilot Study.
Three ranked lists are obtained from the single input RCC dataset.

Pre-QC and post-QC (trim) biomarker lists were then compared to select sets of genes fulfilling the following criteria, also shown in Figure 22:

Set 1: The gene must be ranked highly in the pre-QC list, and ranked much lower on the post-QC (trim) list. This represents a hypothesized *false positive*.

Set 2: The gene must be ranked highly in the post-QC (trim) list, and ranked much lower on the pre-QC list. This represents a hypothesized *false negative*.

Set 3: The gene must be ranked highly in all three lists. This represents a hypothesized *True Positive*.

The genes in each of these sets were then reduced until only five members remained, based on availability of suitable PCR primers. All genes meeting these criteria were then quantified with RT-PCR on a cohort of independent RCC samples. 8 samples each of Clear Cell (CC) and Chromophobe (CHR) Renal Cell Carcinoma (RCC) were used for the study, with cDNA extractions from adjacent slices of each FFPE tissue block constituting duplicate samples. Two tailed t-tests were performed for each gene, comparing the two clinical subtypes. P-values were used to rank each biomarker in order of the degree of successful validation.

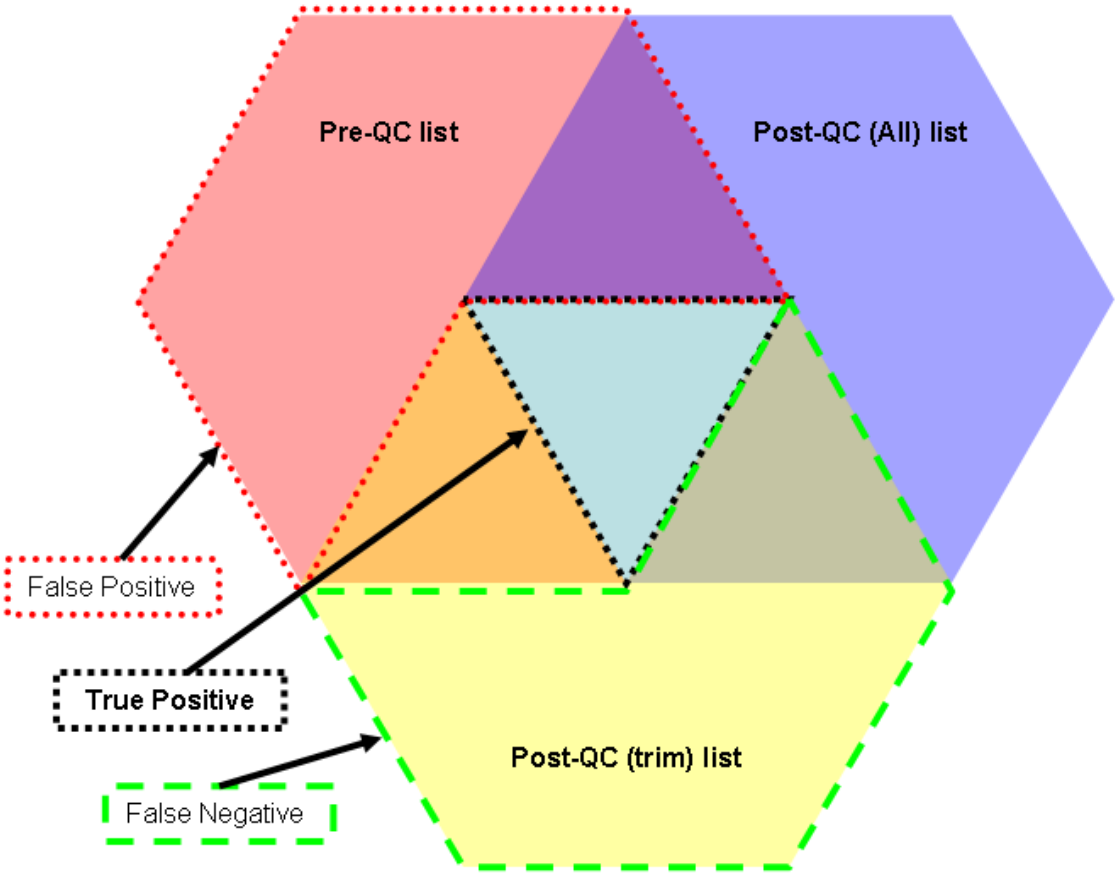

Figure 22: Venn Diagram of QC Predictions
Red, Blue and Yellow solid shading represent the set of top genes in each of the three lists. Outlined areas show the predicted false positive, true positive, and false negative gene subsets for the various combinations of ranked lists.

Forward and reverse primers for each of the 15 target transcripts, and the 18S housekeeping gene were obtained from Applied Biosystems, and quantitative RT-PCR was performed for 40 thermocycles on a 96-well plate. Transcripts that had not appeared after the 40th cycle were recorded as having C_t equal to 40. To reduce the amount of extrapolation and increase signal to noise ratios, samples with average 18S housekeeping gene expression levels higher than 17.5 C_t (indicating low overall cDNA concentration) were thrown out. C_t values for each of the 15 targeted genes were then normalized by subtracting the matched C_t values of the 18S housekeeping gene for each sample.

Results

The results of the pilot study round of PCR validation are summarized in Table 4. Genes in the table have been ranked by p-value, such that genes near the top of the list, particularly the top two, can be said to be 'validated' as biomarkers due to their ability to discriminate between the two RCC subtypes. Each transcript's place in the SVM ranking is shown for three cases in the rightmost columns. First the rank in unprocessed data: *Pre Rank*, Second the ranking after caCORRECT: *Post Rank (All)*, and then the rank after trimming the dataset by removal of the two lowest quality chips: *Post Rank (Trim)*. As can be seen in the table, the there is a trend that the best chance of validation comes from *True Positives*, followed by *False Negatives*, and then lastly *False Positives*. This trend supported our hypothesis but should be expanded for statistical significances and clear impact.

Table 4: Results of PCR Validation.
Genes predicted as true positives are shaded in yellow, False Negatives in green, and False Positives in Red. Ranking by p-value indicates a trend for higher validation rate for true positives, followed by false negatives, and finally false positives.

UNIQID	**Gene Symbol**	Plausibility	**PCR p-value**	**Pre Rank**	**Post Rank (All)**	**Post Rank (Trim)**
202237_at	**NNMT**	Candidate RCC tumor marker J Urol 176:2248 2006	**0.00000**	7	6	9
201835_s_at	**PRKAB1**	energy metabolism	**0.00000**	5	4	7
204396_s_at	**GPRK5**	Signal transduction	**0.00000**	14	11	13
202818_s_at	**TCEB3**	VHL partner	**0.00004**	7153	5049	153
208982_at	**PECAM1**	cell adhesion	**0.00011**	35	160	894
201288_at	**ARHGDIB**	immune and metastasis related	**0.00043**	2	1	3
207042_at	**E2F2**	major gene expression regulator	**0.01725**	1316	235	142
201530_x_at	**EIF4A1**	gene expression	**0.05779**	4	3	4
200853_at	**H2AFZ**	chromatin	**0.07660**	2480	173	161
209451_at	**TANK**	immune related	**0.19730**	1729	106	183
36019_at	**STK19**	MHC gene	**0.21623**	144	2109	1053
211020_at	**GCNT2**	limited	**0.25183**	1041	1095	67
202824_s_at	**TCEB1**	VHL partner	**0.63213**	7609	7316	398
203952_at	**ATF6**	gene regulation	**0.70201**	56	536	1579
209778_at	**TRIP11**	other TRIP's ass'd with hypoxia response	**0.96087**	76	309	1108

Clinical Validation Follow Up Study

Expanding upon the anecdotal evidence suggested by the pilot study, we increased the PCR panel to over 90 targets with the help of a core facility, and tried to expand and reproduce the pilot study which suggested that the use of caCORRECT may increase the reliability of biomarker selection. This time, however synthetic artifacts were added to the original microarray data to enhance the possible magnitude of caCORRECT's impact.

Methods

To determine the effect that caCORRECT had on the ability to correctly identify biomarkers of disease from microarray data, a panel of 96 genes of interest for RCC was assembled for PCR study. These genes were identified from a combination of genes previously identified in the literature as well as a set of genes whose biomarker status was disagreed upon between the caCORRECT and non-caCORRECT versions of the Young et al. data sets (see previous section methods). All PCR analysis was performed on independent patient tissue samples with respect to those used for the microarray analysis.

Gene expression was assessed by quantitative RT-PCR, using total RNA from fixed tissues of 17 clear cell, 13 papillary and 7 chromophobe RCC. PCR was performed with a custom-designed TaqMan Low Density Array (LDA, Applied Biosystems) in a 96-well microfluidic card format, using the ABI PRISM 7900HT Sequence Detection System (high-throughput real-time PCR system). Gene expression data were normalized relative to the geometric mean of two housekeeping genes (18S, ACTB). LDA runs were analyzed by using Relative Quantification (RQ) Manager (Applied Biosystems) software. Relative normalized gene expression was compared in renal tumor subtypes. Genes were declared as being "validated by PCR" if they had an average fold change between classes of magnitude greater than 2.

To measure the effect of caCORRECT in the presence of decreasing data quality, two new versions of the Young et al. dataset [2] were constructed to supplement the original microarray data. Note that our previous study showed no significant spatial overlap between biomarkers and artifacts in this dataset, so any possible benefits of cleaning this particular data was expected to be minimal. Similar to the earlier work done by Jones et al., an even mixture of smaller, less-severe hot-spot and black-hole artifacts were applied to the chips in two sizes (see Figure 23). In order to monitor the effect of artifacts on differential gene finding, as well as the ability of caCORRECT to ameliorate those effects, Receiver Operator Characteristic (ROC) curves were used.

Results

Figure 23 shows examples of synthetic artifacts applied to the Young et al. dataset as visualized by the post-caCORRECT residual images. While the synthetic artifacts may appear more visually stunning than the artifacts "naturally" found in this dataset, they are comparable with those found on other microarrays, such as the one shown covering the left hand side of the chip in Figure 9.

ROC curve analysis of microarray prediction (test prediction) versus PCR validation (ground truth status) is shown in Figure 24. Results show that caCORRECT has little effect on the predictive power of microarrays which are relatively free from artifacts, or contain only weak artifacts. This suggests that the use of caCORRECT is suitable for datasets of unknown quality without serious risk of degrading results from already clean arrays. Furthermore, caCORRECT is able to preserve predictive power of microarrays which are influenced by serious artifacts, which would otherwise suffer from poorer predictive power as measured by area under the ROC curve.

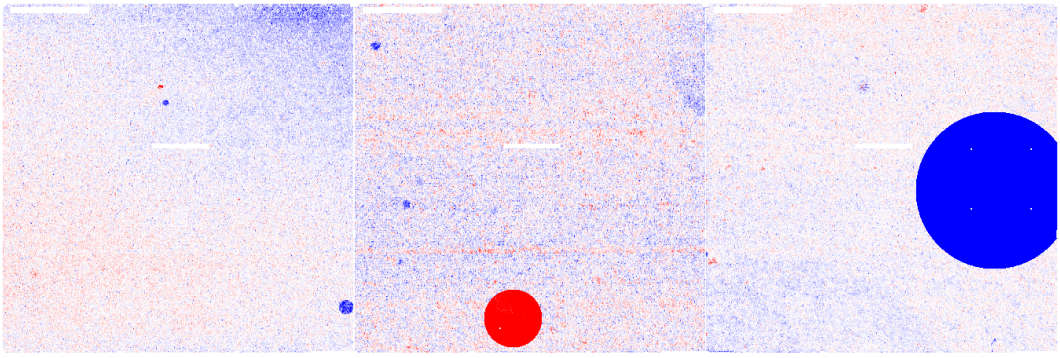

Figure 23: Examples of Artifacts Present for Biomarker Identification Analysis. Images show model residuals after 4 rounds of caCORRECT normalization and artifact identification, but before removal. (left) The most stunning naturally occurring artifacts in the Young et al. dataset, (middle) a synthetic weak hot-spot artifact, and (right) a strong synthetic black-hole artifact.

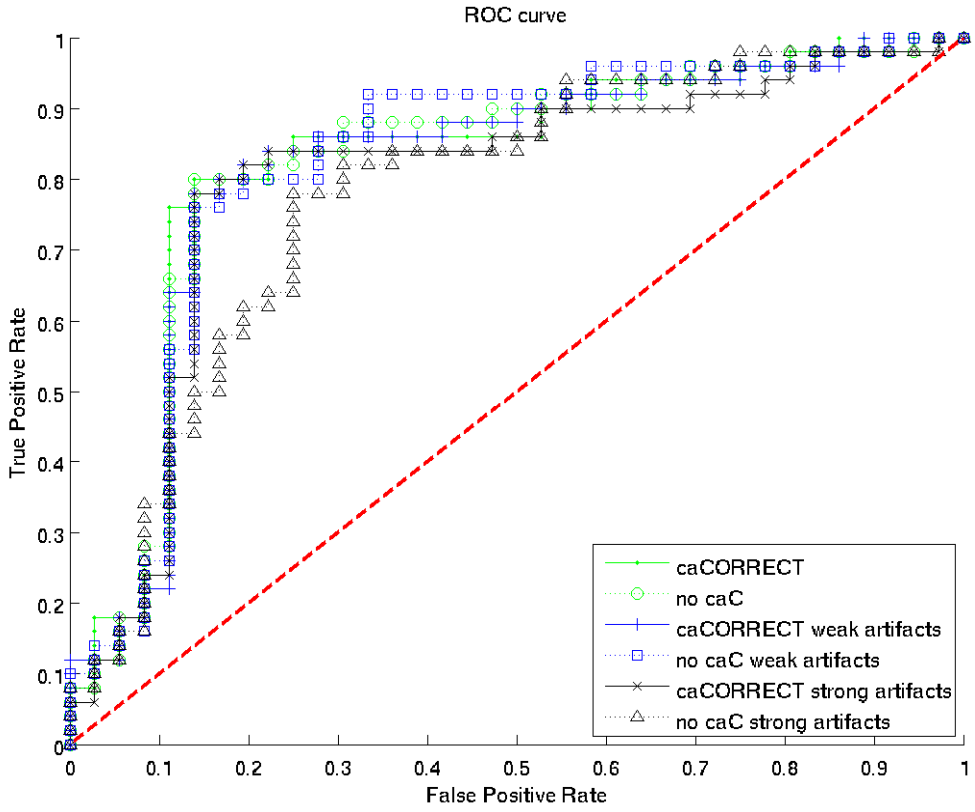

Figure 24: Microarray Fold Change as a Predictor of PCR Fold Change in RCC Samples, and the Effect of Artifacts and caCORRECT Preprocessing.
Genes were thresholded by magnitude of observed log fold change in RMA-derived microarray data, and considered truly differentially expressed if they exhibited more than a 2x or less than a ½x fold change between classes CC and CHR in the PCR data. Only genes for which PCR data were available appear in this analysis. caCORRECT preserves data quality for normal arrays (no loss in the area under ROC curve) and improves quality for arrays which have serious artifacts (recovers lost area under the ROC curve).

Summary

In this chapter, we have used a series of experiments, using a combination of real and synthetic microarray data, to show empirically how caCORRECT can impact the bioinformatics pipeline in various ways. The first test was designed to investigate the impact that artifacts may have had on previously published biomarker results. Results of this test showed that previously published biomarker discovery is sometimes correlated or anti-correlated with the presence of chip artifacts. The second test was designed to see if caCORRECT could improve the poor reproducibility of gene feature selection that is common in microarray studies. Results of this test show that caCORRECT provided a moderate advantage in the reproducibility of ranked biomarker lists during cross-validation. The third test was designed to directly quantify the effect that caCORRECT has on the accuracy of gene expression in the presence of artifacts. Results show that caCORRECT reduced the RMS error of existing gene expression methodologies by anywhere from 20-66% depending on the nature of the artifacts and the 3rd party gene expression algorithm being evaluated. After establishing that caCORRECT increases accuracy of gene expression, we then tested to determine the effect that caCORRECT has on the sensitivity and specificity of biomarker selection, and found that caCORRECT provides a moderate advantage when the raw microarray data contain serious artifacts. In this way, biomarkers selected using caCORRECT are more reliable, and have a better chance of validation on external samples.

CHAPTER 4

DEVELOPMENT OF A BIOMARKER BASED DIAGNOSIS

The final deliverable for this dissertation is a clinically relevant molecular screening for an unknown subtype of RCC. Renal Cell Carcinoma (RCC) is used as a case study primarily because relatively little research has been done for the molecular classification of RCC. Moreover, the subtyping problem for RCC is more difficult than simple normal versus cancerous tissue classification, and should prove to be more useful in a clinical setting. RCC is the most common form of kidney cancer arising from the renal tubule in adults [96], and more than 90% of clinically significant lesions can be diagnosed as one of the common subtypes of renal tumor: clear cell RCC (70-75%), papillary RCC (10-15%), chromophobe RCC (2-5%), or renal oncocytoma (5%). Renal tumor subtypes exhibit several common morphological characteristics, making diagnosis difficult and subjective in many cases [85]. Quantitative molecular classification is therefore promising as an alternative or supplement to morphological classification for the diagnosis of RCC. Proper classification of RCC is important, because each of its subtypes is associated with a distinct clinical behavior, requiring different treatment courses.

RCC Biomarker Selection and PCR Validation

Gene biomarker selection was done as described in the earlier sections describing the clinical validation of caCORRECT. As part of this validation of caCORRECT, 15 transcripts were identified (see Table 4 for a full list) for quantification by PCR. Two biomarkers in particular, NNMT and PRKAB1, performed exceptionally well both in the original microarray data (Figure 25), and during the first set of independent clinical samples (Figure 26). Using 18S-normalized expression for PRKAB1 and NNMT alone, 100% classification was possible. A follow-up round of PCR on a new cohort of 9

chromophobe vs. 17 clear cell vs. 13 papillary was also performed, the results of which are also shown in Figure 26. Although a batch-effect can clearly be observed between the two rounds of PCR, it is still possible to design a classifier that achieves 100% separation between all clear cell and chromophobe samples. This separation is especially encouraging considering the subtle changes in protocol between the original study, performed by myself, Qiqin Yin-Goen, and James Torrance, and the follow up study, performed by a professional core facility.

Of the 90 markers identified by the RCC microarray experimentation, over 10 have been verified as potentially excellent markers by follow-up PCR measurements. So far, only PRKAB and NNMT have been assayed for protein expression.

Depending on time constraints and the availability of data, this entire translational biomedical informatics methodology could be applied to subtyping problems other than Clear Cell versus Chromophobe, however currently PCR validation data is only available in sufficient quantities for these subtypes. Some data from Papillary RCC samples have also been assayed, although the number of successful PCR-validated Papillary versus other RCC markers is understandably lower because the genes were not selected under this premise.

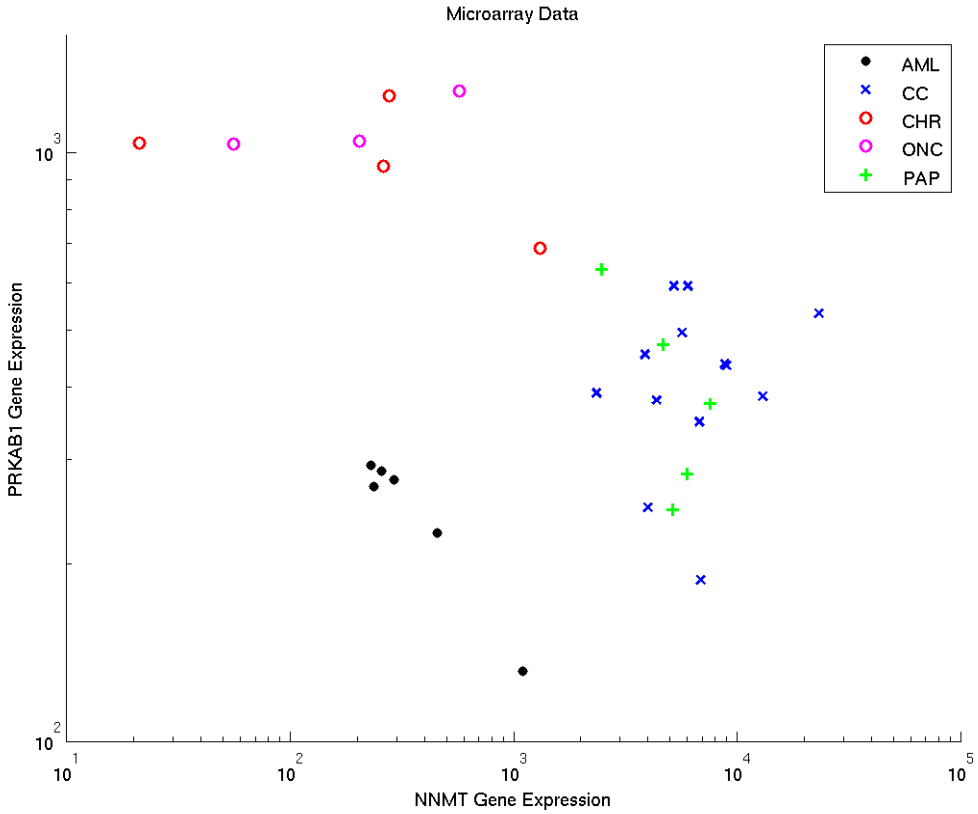

Figure 25: Microarray Gene Expression of NNMT and PRKAB1 in RCC Tissue. Gene expression is in arbitrary units on the log scale.

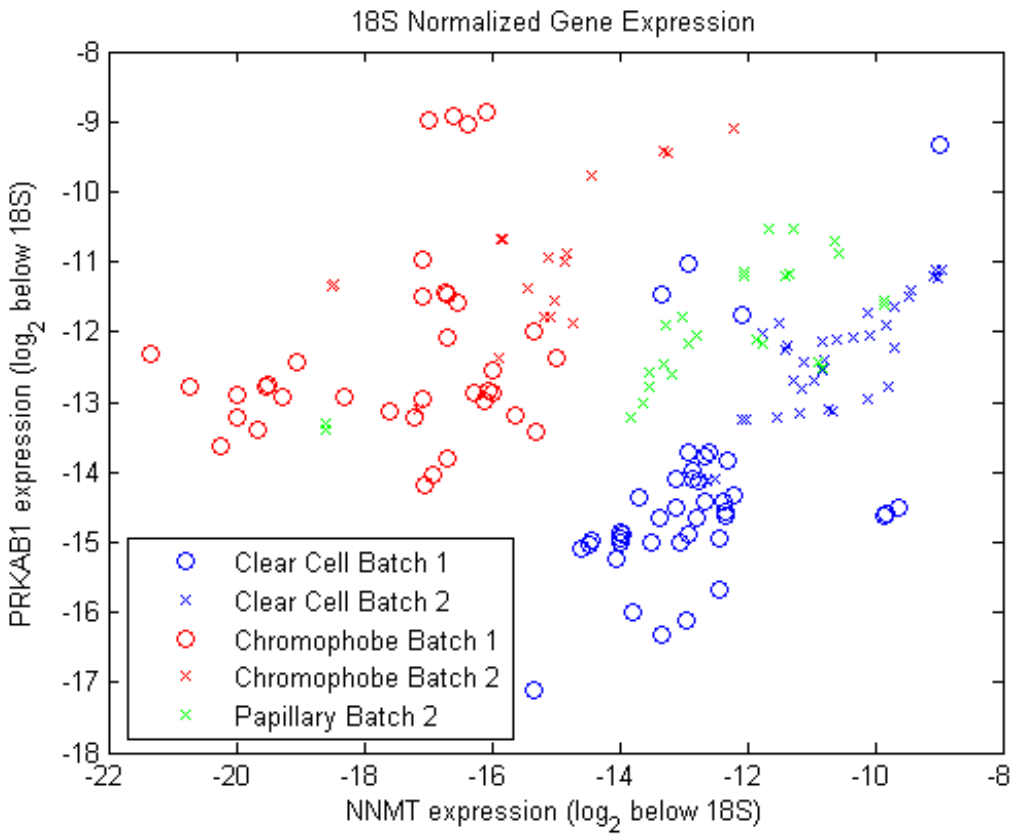

Figure 26: Self-normalized Gene Expression of Biomarkers NNMT and PRKAB1 in RCC Tissue.
Gene expression is shown in the log base 2 scale, consistent with the round of identification during PCR amplification. Each sample was normalized with an internal control by subtracting the value by the 18S value (division in the linear scale).

Quantitative Protein Expression Analysis

Because of the good results of gene based classification, NNMT and PRKAB1 were chosen to be the first candidates for multiplexed quantum dot analysis. Multiplexed QD immunohistochemistry data were obtained from a Biomax BC07015 RCC tissue microarray stained with two antibody-conjugated Quantum Dot (QD) solutions according to the protocol in Xing et al.[75].

Images were taken from each tissue sample using the Olympus IX71 microscope equipped with a Nuance multi spectral imager. In an attempt to minimize confounding variables which may undermine quantification, each image was taken with the same objective lens as well as with the same exposure time during a single session at the microscope (~2 hours total time). The fluorescence of each sample was recorded at wavelengths between 500 and 800nm in 10nm increments. This range was chosen because it corresponds to the active fluorescence emission range of the QDs.

Quantification was done as previously described in the work by Caldwell et al.[16]. Briefly, a positively-constrained least-squares unmixing procedure was used for spectral unmixing of the two QD signatures and two known RCC autofluorescence signatures, followed by an average pixel intensity calculation for each unmixed QD component across the entire area captured by the microscope camera setup. The resulting data are semi-quantitative in the sense that they may justifiably be compared to each other quantitatively, but are not necessarily suited for direct comparison to new samples acquired on different microscope setups or stained with different lots of reagent.

Initial Results

As an initial attempt at quantitative analysis, QD-probes for the novel markers PRKAB1 and NNMT were multiplexed along with QD-probes for the known cancer marker MDM2 and a control marker β-Actin. A constrained least-squares method was

used to accomplish the spectral unmixing of the PRKAB1 and NNMT intensities from the other two QD-antibody conjugates used in the stain. From these unmixed intensities, pseudo colored images were produced showing both biomarkers of interest in contrasting color (see Figure 27). The global intensities of each QD were then calculated with a crude global average intensity to assess potential as features for classification (see Figure 28). Using a simple Bayesian classification rule under Gaussian assumption, 28/32 = 87.5% clear cell samples and 16/18 = 88.9% chromophobe samples are classified correctly.

During parallel work as part of an investigation into the reproducibility and quantification potential of QD-IHC, it was discovered that some of these findings may not be directly due to protein expression. Instead, it is hypothesized that much of the observed NNMT signal is due to tissue autofluorescence, and imprecise unmixing. Furthermore, other experiments suggest that there may be significant crosstalk between probes which may also explain these results. Discussion of these issues is the subject of the next chapter. Regardless of the true nature of the signal, be it autofluorescence or protein biomarker, the technology does produce usable results as-is. In other words, even if the horizontal axis label of Figure 28 is inaccurate, the classification accuracy is undeniable. Even so, in an effort to do good science, this experiment was repeated with a modified protocol that minimizes crosstalk between probes and confounding from autofluorescence.

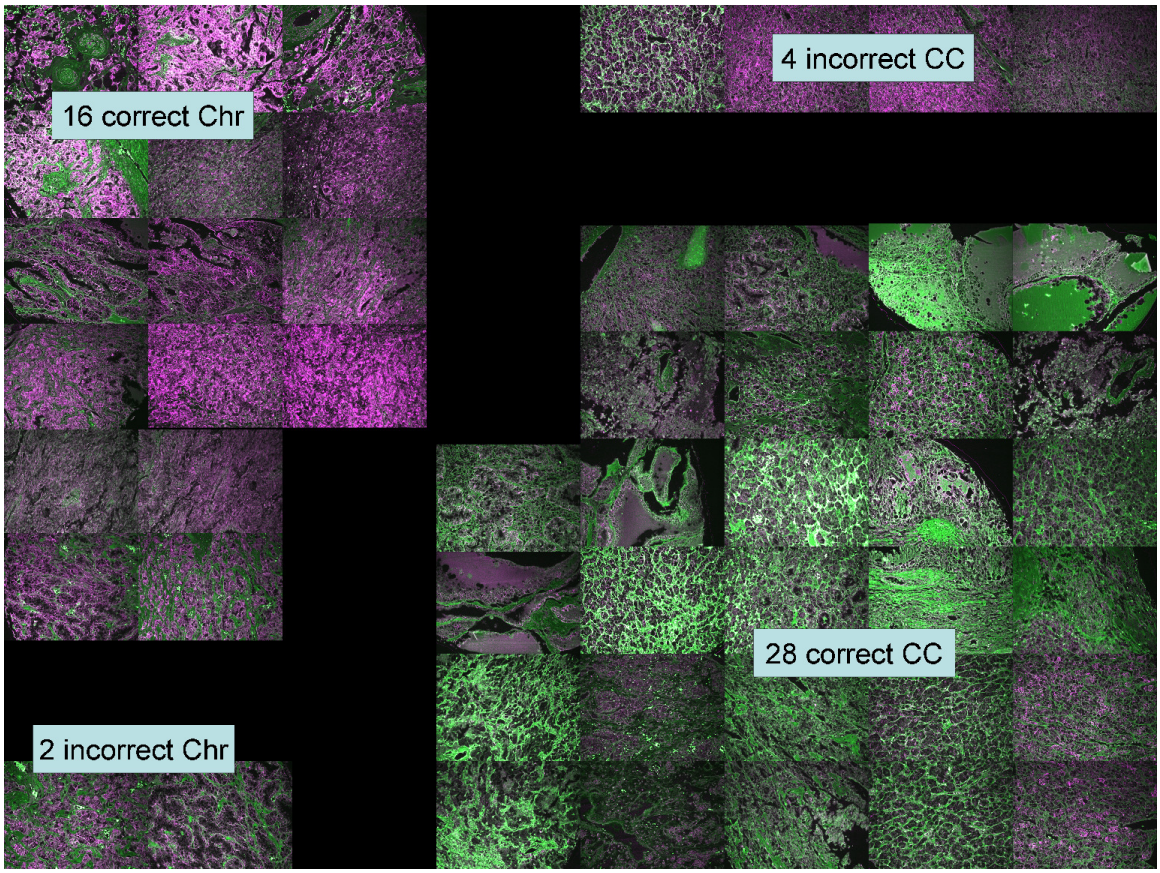

Figure 27: Pseudocolored Images of Quantum Dot Staining of RCC Tissue Microarray Samples.
PRKAB1 staining is shown in magenta, NNMT in green. Due to imperfect unmixing, NNMT signal may also contain autofluorescence signal. Sample staining is courtesy of Jian Liu.

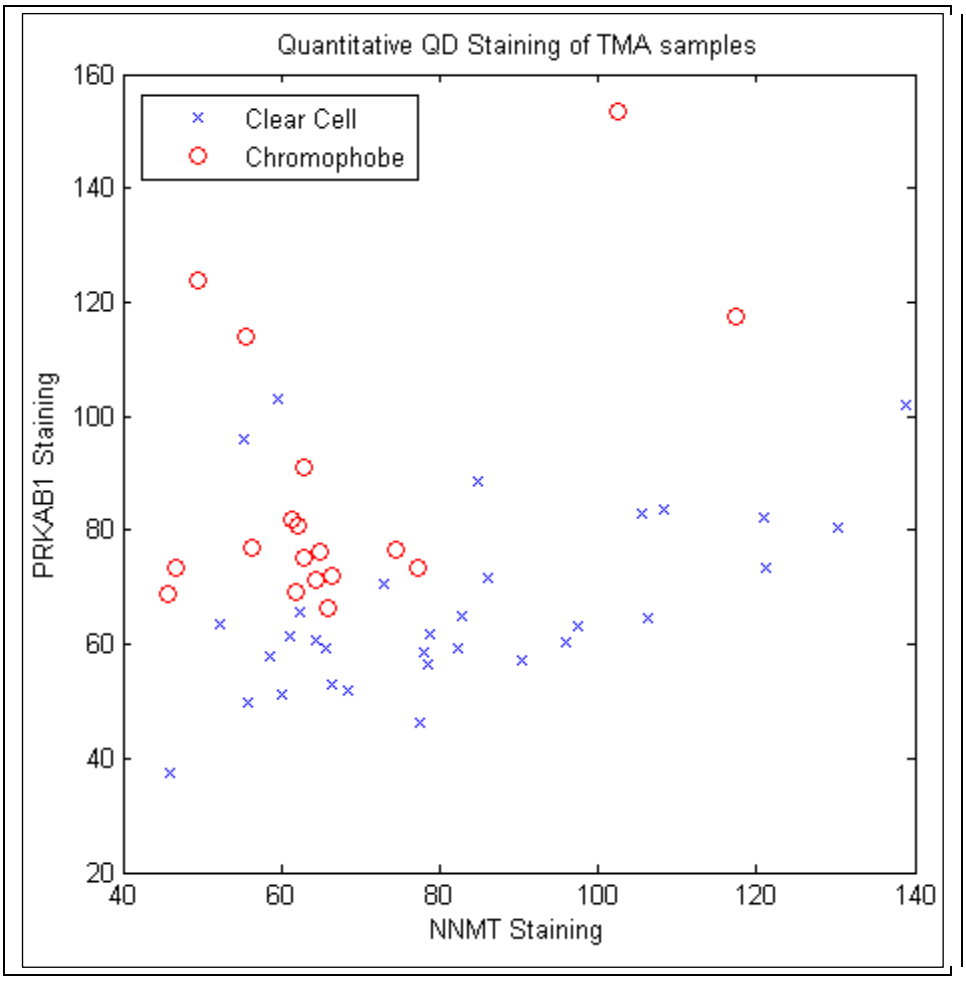

Figure 28: Quantum Dot Staining of RCC Tissue Microarray Samples.
Staining is measured in arbitrary units which are comparable across both PRKAB1 and NNMT staining. Sample staining is courtesy of Jian Liu.

Revised Results

In a revised protocol, Quantum Dot antibody conjugates were constructed such that one QD with emission peak at 605nm stained for PRKAB1, and another QD with emission peak at 655nm stained for NNMT. Unlike the previous experiment which reused secondary antibodies and stained serially, this experiment uses completely independent secondary antibodies, and a parallel staining protocol to minimize crosstalk. Furthermore, the revised protocol makes use of the two available QDs which have the largest signal to noise ratios with respect to tissue autofluorescence at peak emission.

Protein analysis once again confirmed the trends observed in gene-based analysis, but with less reliable separation than either the PCR or the previous attempt at protein quantification. Figure 29 shows examples of pseudocolored photomicrographs for a CC and a CHR sample, as viewed with a custom-built interactive GUI. Automated results of quantification, not including any region of interest selection, are shown in Figure 30. Investigation of photomicrographs reveals that, while overall tissue expression levels of NNMT were very low (< 5 Signal to Noise ratio, and <50% of autofluorescence signal), vascularized tissues and red blood cells were detected as expressing small amounts of NNMT. This suggests that the elevated levels of NNMT observed in CC samples in this study could also be due to the well known phenomenon of increased “chicken-wire” vascular patterns that are a hallmark of CC samples. Further investigation shows that many CHR samples which express NNMT highly were either exceptionally abundant with red blood cells in the field of view, or near large blood vessels.

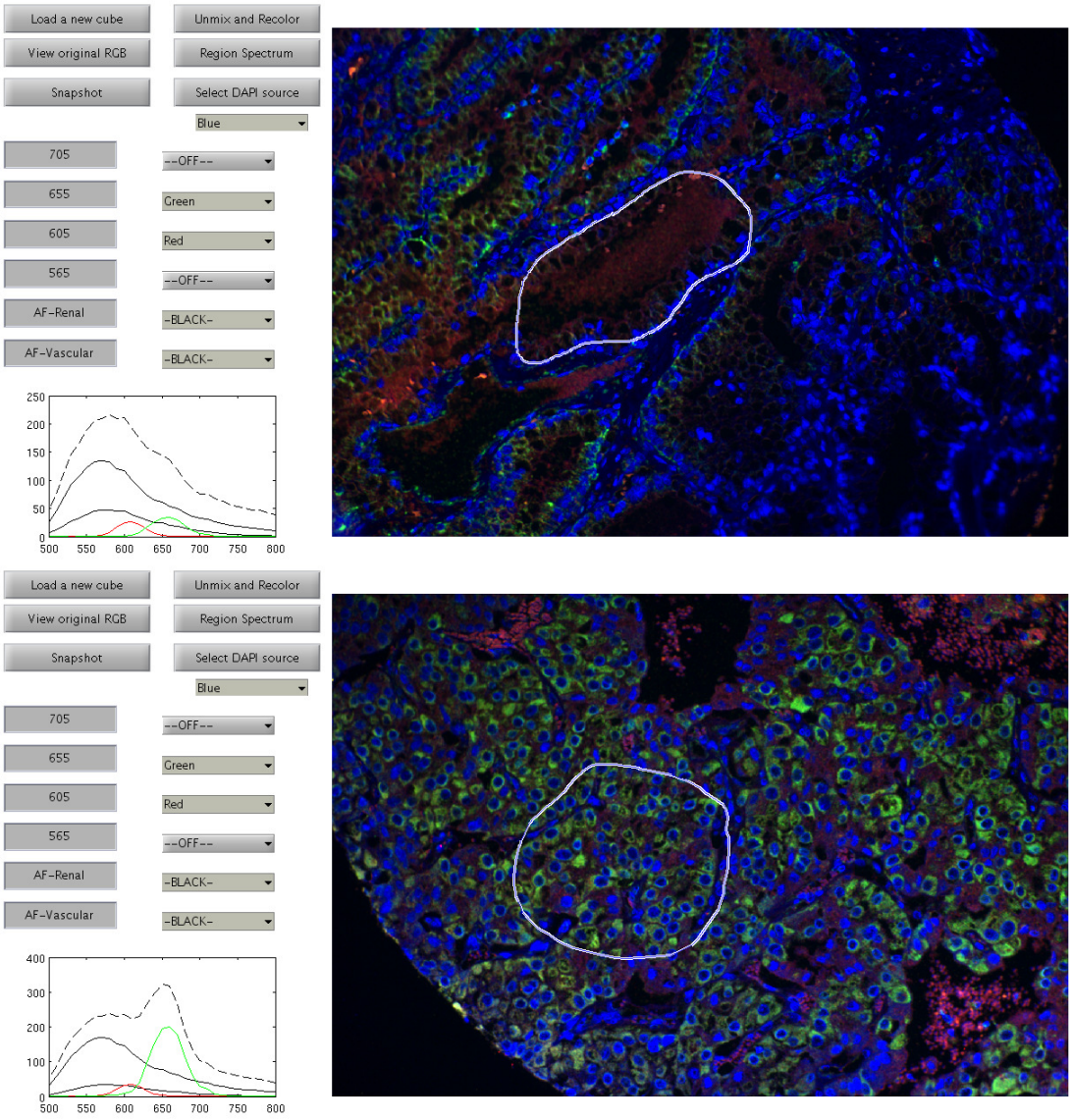

Figure 29: Pseudocolored Images of Quantum Dot Staining of RCC Tissue Microarray Samples.
Screenshots from our custom quantification GUI are shown for a clear cell sample (top) and chromophobe sample (bottom). PRKAB1 staining (QD605) is shown in red, NNMT (655) in green, and DAPI counterstain in blue. Regions of interest are shown outlined in white, with corresponding color coded spectral components shown in the bottom left corner of each panel. Large autofluorescence components are visible in both spectra plots, but the chromophobe sample has higher expression of the PRKAB1 biomarker. Sample staining is courtesy of Jian Liu. Imaging is courtesy of Matthew Caldwell.

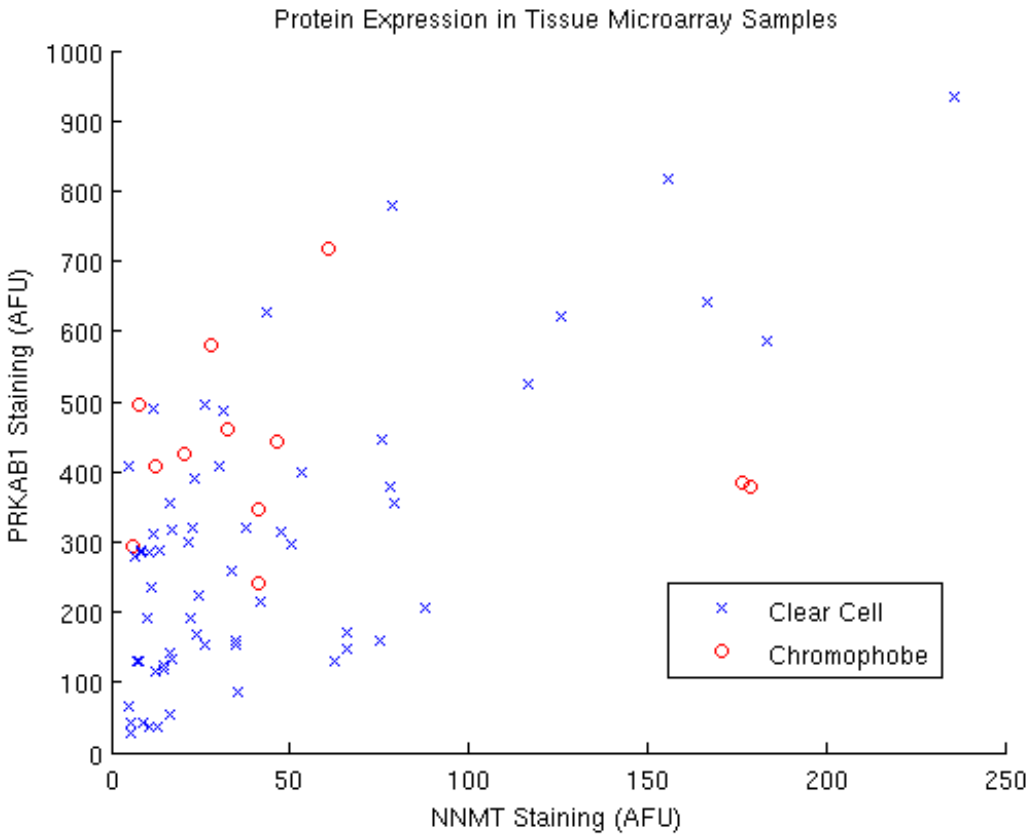

Figure 30: Quantum Dot Staining of RCC Tissue Microarray Samples.
Staining is measured in arbitrary units, but is quantitative in nature. The trend of expression suggested by RNA experimentation is still evident, although it is less clear, possibly due to the low NNMT absolute signal levels observed in all samples. Sample staining is courtesy of Jian Liu. Image acquisition is courtesy of Matthew Caldwell.

Summary

In Chapter 4 we have produced the first clinically relevant result attained with the help of caCORRECT. Using a case study of RCC clinical samples, biomarkers were first selected from microarray data, and then validated with quantitative RT-PCR and quantum dot immunohistochemistry (QD-IHC). We have shown that a RT-PCR panel of two markers, NNMT and PRKAB1, can be used in tandem to create an extremely high accuracy (100%, n = 24) RT-PCR based classification system for clear cell versus chromophobe RCC. We then show the results of two subsequent QD-IHC studies. The first study suggests accuracy as high as 88% can be achieved, but these results were not entirely reproducible, most likely due to a combination of high background autofluorescence and low signal put out by the QDs chosen for the experiment. The second, follow-up experiment had a higher signal to noise ratio but produced a less-clear classification result. A curious consequence of these two experiments suggests that magnitude of tissue autofluorescence may actually prove to be a viable marker for the differentiation of RCC tumor subtypes. Issues which lead to this low signal to noise ratio and low reproducibility in QD-IHC are discussed in the next chapter.

CHAPTER 5

TOWARDS A QUANTITATIVE QUANTUM DOT METHODOLOGY

The key challenge for quantitative multiplexed QD-IHC is the development of a protocol which is reproducible and reliable across technicians and laboratories. Similarly to the story with microarrays, without attaining such reproducibility, QD-IHC cannot achieve maximum impact in a clinical environment. Development of clinically viable method of QD-IHC represents the last major frontier to be conquered in the translational bioinformatics pipeline. Compared to microarray QC, QD QC is an underexplored field. This chapter outlines current issues which hamper reliability of QD-IHC, and discusses recent progress towards overcoming these challenges. Methods of spectral unmixing and cross-reactivity have been described in my previous publications in EMBC conference publications in 2008 [16] and 2009 [15], respectively.

Spectral Unmixing Model

Spectral unmixing in the context of QD-IHC refers to the process of source separation for an observed wavelength-resolved spectrum obtained from an imaging device. For a typical QD-IHC experiment, sources are expected to be both the quantum dots used for staining as well as one or more autofluorescence signatures occurring naturally in tissue. Sample spectra are obtained by monitoring the fluorescence intensity of a stained tissue sample at multiple wavelengths in the visible and near infra-red range. For our microscope setup, intensity is measured at 31 equally spaced wavelengths from 500 to 800nm.

To spectrally unmix an entire image, one must unmix each pixel in the image. For simplicity of notation, but without loss of generality, we will consider only the case of a single pixel. The spectrum of a single pixel can be represented in a column vector, $\mathbf{y} \in \Re^{31}$. Likewise, the j^{th} isolated source spectrum of a single quantum dot or

autofluorescence can also be can be represented as the j^{th} column in a matrix, **A**, with dimension 31 by N (N being the total number of fluorescence sources). The observed spectrum, **y** can then be modeled by the following linear system of equations:

$$\mathbf{y} = \mathbf{Ax} \tag{5.1}$$

where $\mathbf{x} \in \Re^N$ is a column vector such that the j^{th} element of **x** is the magnitude of the contribution of the j^{th} source to the pixel's spectrum. This unmixing model for fluorescence is purely additive such that the observed intensity spectrum of each pixel is composed of the summation of each individual QD contribution plus autofluorescence. Note that this model will break down when the data acquisition is saturated, so care must be taken not to overexpose the image. Assuming that this additive fluorescence model is true, and that the spectra of all sources are linearly independent, 31 wavelength-resolved measurements help construct an over specified system of linear equations which theoretically provides more than enough information to resolve even a dozen QD signals. In practice, however, it is rare to use more than 6 QDs at a time. Spectral unmixing is the process of estimating **x** when given only **A** and **y**. We estimate **x** by minimizing the mean squared error between **y** and **Ax**, with the constraint that the elements of **x** be non-negative. This constraint ensures that all sources contribute positively to the observed fluorescence; in other words, QDs cannot be found to have removed light from the system.

Problems arise in the unmixing problem whenever assumptions are violated. The first standard assumption for solving these types of equations is that the error term, **ε**, is distributed with zero mean. The form of the equation which includes the error term is

$$\mathbf{y} = \mathbf{Ax} + \boldsymbol{\varepsilon} \tag{5.2}$$

To test this assumption, a series of pure QD in solution were monitored for baseline noise. For out particular imaging setup, it was found that **ε** is distributed with a median of 26 (Figure 31). Interestingly, this value did not scale with exposure time. With

this in mind, we reformulate the mixing equation to include baseline subtraction for each element of **y**. For our hardware setup, $\mathbf{b} \in \Re^{31}$ and has each element equal to 26.

$$(\mathbf{y} - \mathbf{b}) = \mathbf{Ax} + \boldsymbol{\varepsilon} \tag{5.3}$$

It is expected that the distribution of baseline noise, and thus the values of **b**, will change from hardware to hardware. As such, each setup should be carefully tested and calibrated to ensure maximum accuracy.

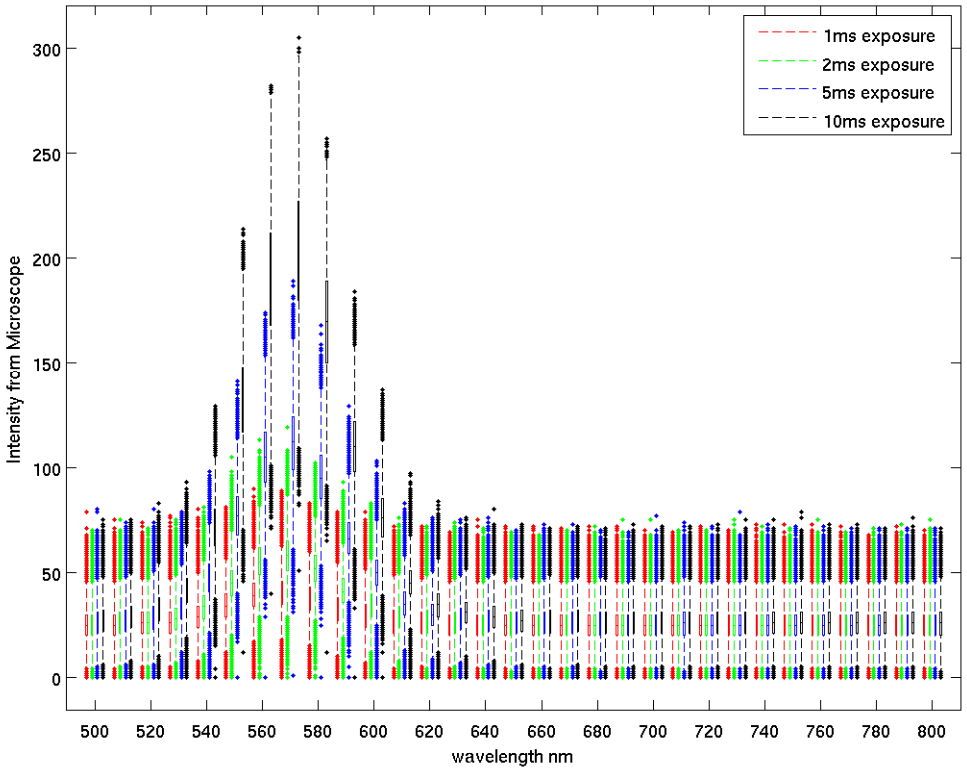

Figure 31: Box plots of Fluorescent Intensity versus Wavelength for Pure QD in Solution at Different Exposures.
Images are of a drop of pure solution of QD565 in water, taken at increasing exposure times. Box and whiskers plots are created from the multiple pixels present in a single image. Outside of the emission range for the QD, a baseline with mean 26 can be observed.

Characterization of QD and Tissue Autofluorescence Spectra

Integral to the solution of the unmixing equation is knowledge of the matrix, **A**. Even if one has perfect knowledge of the columns of **A,** corresponding to QDs based on manufacturer specifications or other similar analysis, the columns of **A** which describe autofluorescence components must still be derived empirically. Furthermore, we have found that manufacturer's specifications for QD emission spectra cannot be trusted to be reproduced by our own multispectral imaging setup (Figure 32). Using the variable substitution ($\mathbf{y} - \mathbf{b}$) = $\tilde{\mathbf{y}}$, the unmixing model can once again be written in the linear form:

$$\tilde{\mathbf{y}} = \mathbf{A}\mathbf{x} \tag{5.4}$$

A single multispectral image is composed of hundreds of thousands of pixels which constitute a set of many different $\tilde{\mathbf{y}}$. Stacking many $\tilde{\mathbf{y}}$ side by side to form a matrix **Y**, and stacking many **x** side by side to form a matrix **X**, allows another re-write of the unmixing equation with the new feature that the rows of **X** may be interpreted as images, each of which is a single fluorescent component's contribution to the overall image.

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \tag{5.5}$$

This form of the equation is familiar to the source separation community, and many standard algorithms exist to learn **A** and **X** simultaneously from **Y**. Perhaps the most widely understood method is the eigen-decomposition of **Y**, otherwise known as Principal Component Analysis (PCA). In a multispectral image taken of only a single QD in solution, we would expect the first principal component of **Y** to be a good estimate of the column of **A** describing the spectrum of that QD, and this is indeed the case. If two QDs are mixed homogenously in solution, or if their protein targets are co-localized in a tissue sample, the case is less clear. In these cases, the two QD expressions are expected to be highly correlated with each other, and thus PCA will most likely result in the largest

spectral component having elements of both QDs in it. Such a result would be inappropriate for inclusion in the **A** matrix. For this reason, automated estimation of the columns of **A** should begin with either only pure, single component, images, or with diverse, heterogeneous images.

One problem with PCA is that it can produce spectral components with negative elements. In the context of fluorescence imaging, such a negative intensity is nonsensical and should be avoided. Non-Negative Matrix Factorization (NNMF) provides an attractive alternative to PCA in these circumstances, but comes with two caveats: NNMF is usually more computationally complex than PCA, and NNMF must be seeded in order to ensure reproducibility. A simple trick that we use to ensure reproducibility is to seed the NNMF with the absolute values of the results of PCA. Figure 32 shows a case study of the results of learning QD spectra from isolated pure solutions using PCA and NNMF versus the spectra supplied by the manufacturer of the QDs (Invitrogen).

As a further validation of use of the learned NNMF spectra, an image of a solution of four QDs were unmixed using both the manufacturer's spectra, and those learned from NNMF as shown in Figure 32. While unmixing is the process of learning **x** from **A** and **y**, reconstruction is the reverse process, creating a model of **y** from **x** and **A**. Figure 33 shows the result of the reconstructed spectra using the manufacturer's spectra, and Figure 34 shows the result when using spectra learned from NNMF.

Unlike QD spectra, tissue autofluorescence source spectra cannot be easily learned by isolating components in a solution. Autofluorescence for tissue is instead learned by imaging tissues that have been stained with every QD-IHC reagent except for the actual QDs. NNMF analysis of breast tissue, for example, yields two distinct components of autofluorescence. Unmixing of a multispectral image of unstained breast cancer tissue using these two spectra reveals two distinct tissue types which can be differentiated based on autofluorescence, as seen in Figure 35.

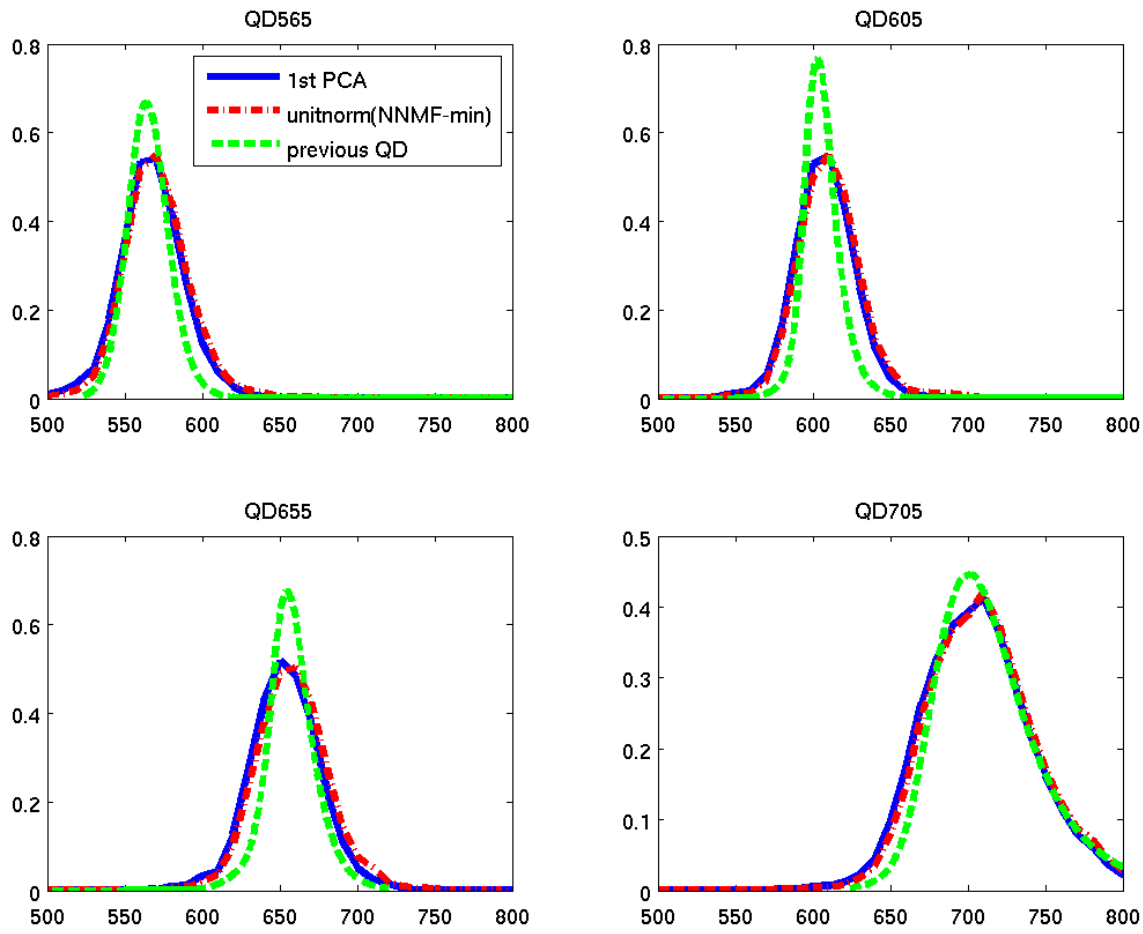

Figure 32: Comparison of Manufacturer Provided Spectra, PCA Spectra, and NNMF Spectra.
All spectra have been normalized to have unitary Euclidean magnitude. Both of the spectra which are derived empirically are visibly different from those supplied by the manufacturer

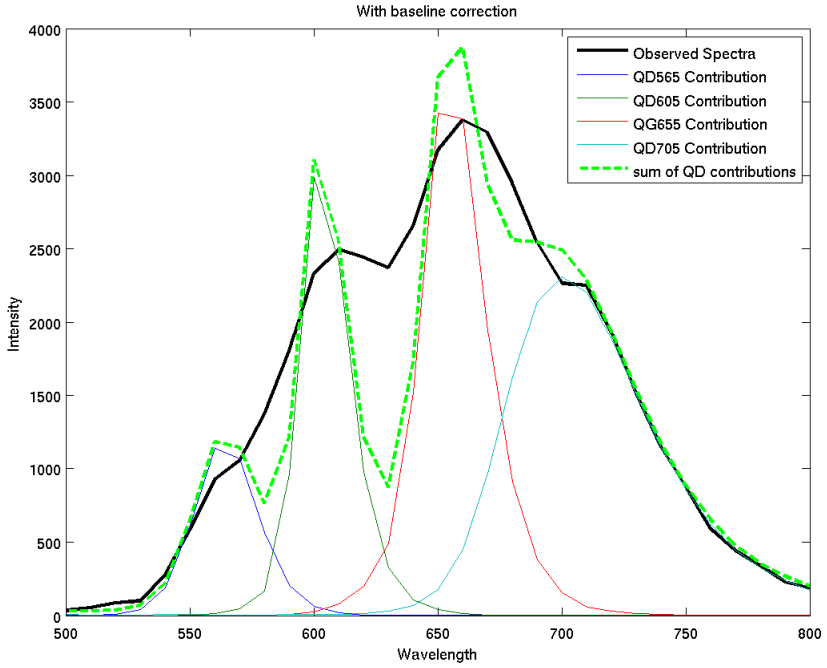

Figure 33: Unmixing Result Using Manufacturer's Spectra.

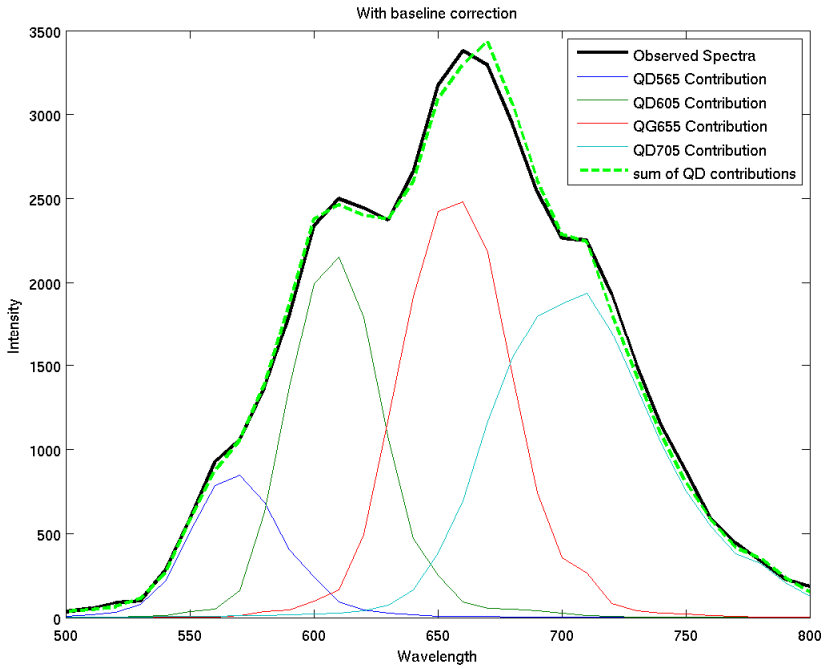

Figure 34: Unmixing Result Using Spectra Learned from NNMF.

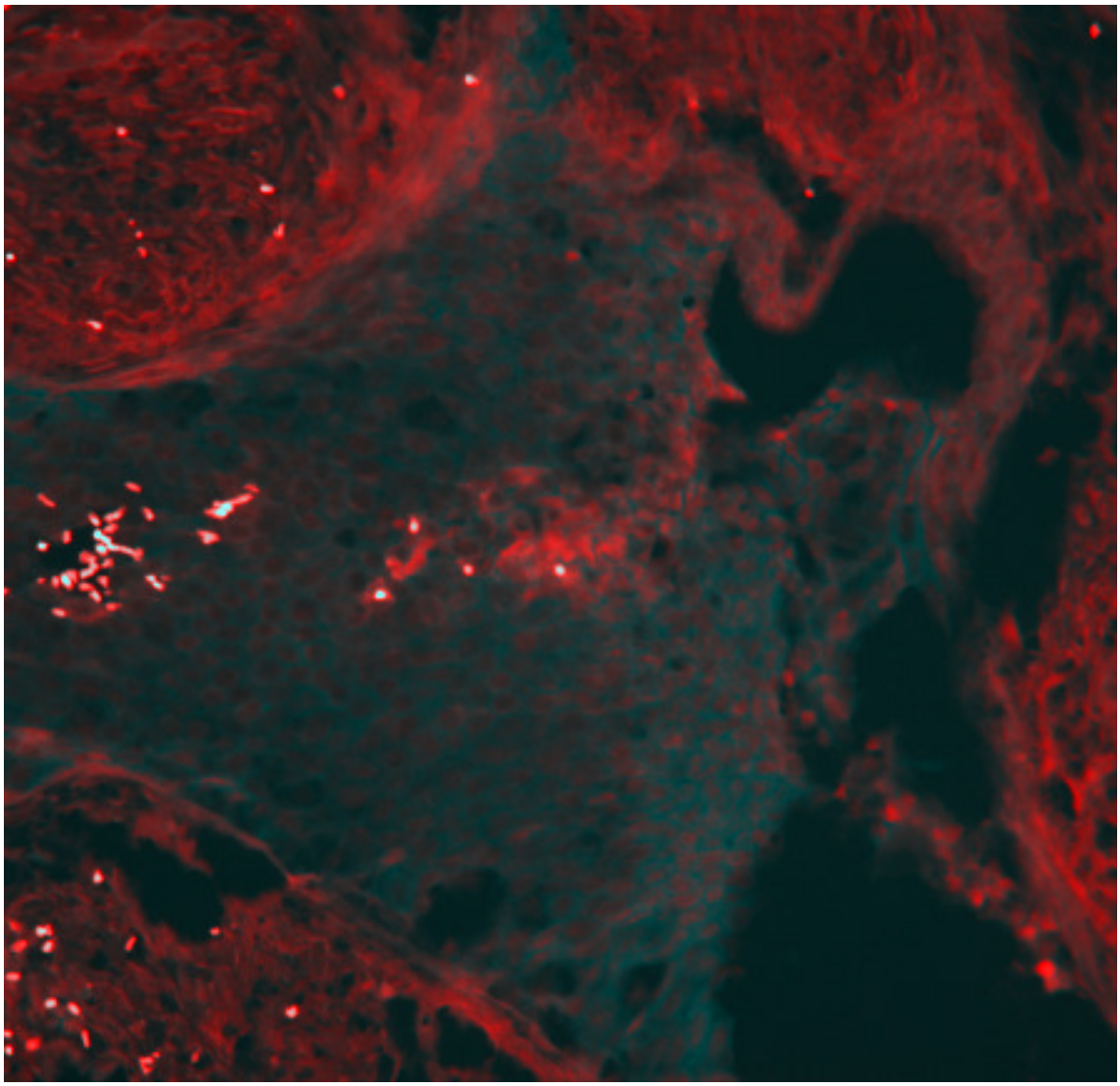

Figure 35: Pseudocolored Image Showing Two Regions of Tissue Autofluorescence in Breast Tissue.
Epithelial tissue shows type 1 autofluorescence (red), glandular tissue shows type 2 autofluorescence (cyan), and red blood cells exhibit a mixture of both.

Differences between QDs Affecting Interlab Comparison

Other factors which may affect the accuracy of absolute QD-IHC image quantification include the relative brightness of different quantum dots, and differences in the number of antibody binding events per QD. To ensure that quantification is independent of QD, standard curves must be created for each QD which describes the effect that the type of QD has on the quantification of protein. The results shown in Figure 33, for example show a mixture of QDs 565:605:655:706 in the ratio of 5:2:1:5. Even in these corrective ratios, the highest concentration QD (QD565) is still less intense than the most dilute QD (QD655). This problem of weak signal for lower wavelength QDs is aggravated by their overlap with common tissue autofluorescent signatures (see Figure 36).

The magnitudes of these unmixed components are still in arbitrary units due to the uncharacterized relationship between affinity for antigen and fluorescent intensity per quantum dot. While these factors may be very difficult to account for, a clinically viable assessment of staining is still possible if reagents can be mass-produced, even if the units are arbitrary. To achieve such repeatability, all staining reagents must be used in abundance, and with sufficient incubation times to ensure that target proteins are the limiting factor in binding. Based on calibration data which specify the relative brightness of QD probes (such as shown in Figure 33), unmixed QD intensities may be scaled to reveal a more universally comparable measure of protein expression. Furthermore, comparing these intensities to standards, such as a fixed amount of pure QD in solution can then lead to repeatability across hardware setups.

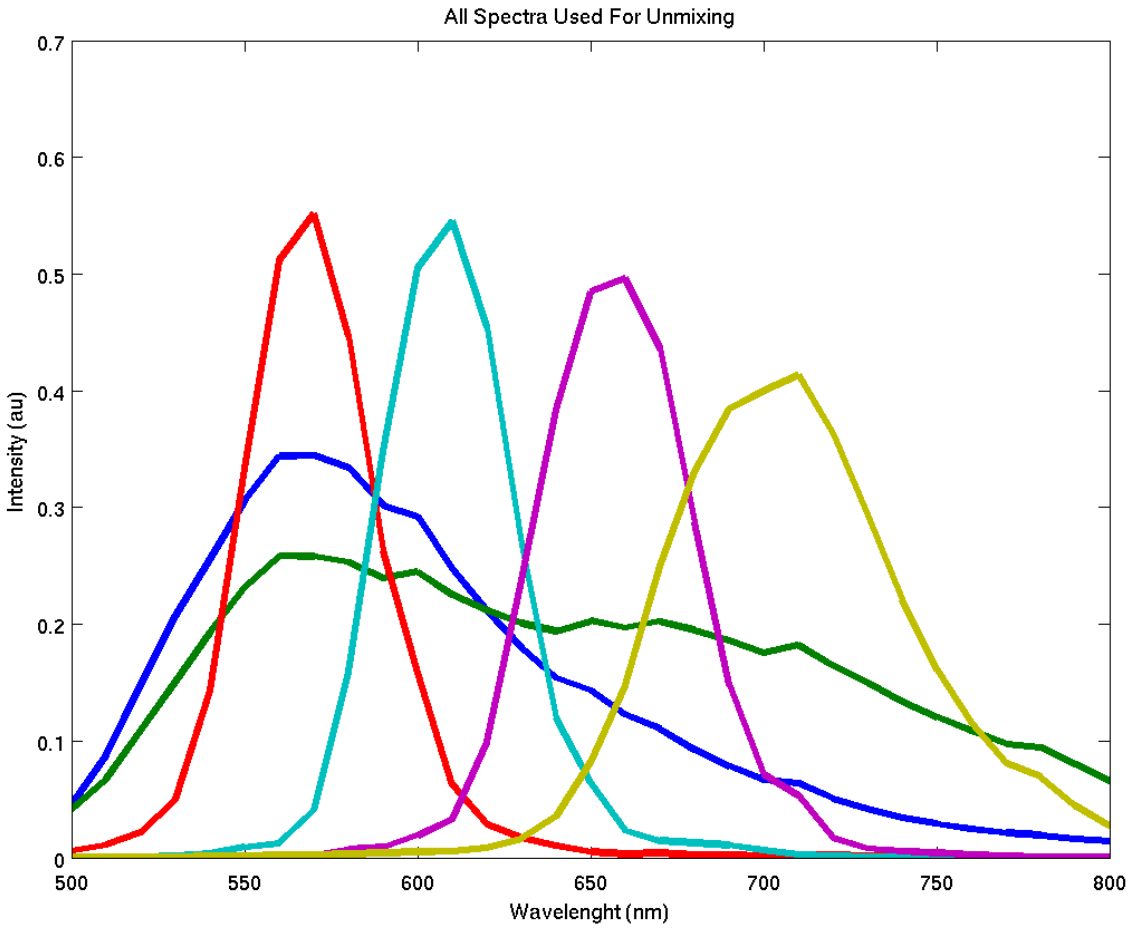

Figure 36: Learned Spectra of 4QDs and 2 Autofluorescent Components.
QD spectra are colored red, cyan, magenta, and gold, and autofluorescent spectra are colored blue and green. All of these 31-dimensional spectra have been normalized to have Euclidian norms of 1. The emission peaks for QD 565 and both autofluorescent signatures overlap.

Antibody Crosstalk during Multiplexing

I have previously proposed a method to account for this antibody cross-reactivity after imaging is complete [15]. Before quantification, each image is segmented to include only regions of interest according to the original intended purpose of the image. For example, breast cancer images which are stained for progesterone receptor are segmented to include only glandular regions of the tissue where the protein is expected. Following this segmenting process, spectral unmixing is performed as usual to produce one image for each QD source intensity. 2D histograms are then constructed from the intensities of corresponding pixels from the intensity images of the first and second QD which share the same secondary antibody targets (Figure 37, top panel). Histograms are constructed to contain 256 equal-sized bins from zero to the maximum image intensity.

For each histogram bin of the first QD image that contains at least 20 elements, the bottom 20th percentile of the second QD's intensity is estimated (Figure 37, top panel, green line). From the paired list (first QD intensity, corresponding second QD's lower quintile), the centroids of the first and last three pairs are calculated. The line which connects these two centroids is called the crosstalk estimate line, and serves as a linear model of interaction between presence of primary antibody (as assessed by the first QD signal) and improper presence of secondary QD due to cross antibody reactivity (Figure 37, top panel, blue line).

The slope and intercept of the crosstalk estimate line are calculated such that the line may be represented by the following equation:

$$QD2\_crosstalk = b + m*QD1 \tag{5.6}$$

The observed signal for the second QD is assumed to be a combination of true signal and crosstalk according to the following model:

$$QD2\_observed = QD2\_crosstalk + QD2\_true \tag{5.7}$$

Thus, a better estimate of the true signal of the second QD can be constructed by subtracting the expected value of the crosstalk-component (calculated with the first QD signal) from the observed second QD signal.

$$QD2\_true = QD2\_observed - (b + m*QD1) \quad (5.8)$$

Using this correction equation, the intensity image of the second QD (Figure 37, bottom left panel, green color) can be modified on a pixel-by-pixel basis using the information from the first QD intensity image to obtain an approximation of the original image which has been adjusted for crosstalk (Figure 37, bottom right panel, green color). These images may then be quantified using standard procedures. In the example image, the green-colored QD appeared to be co-localized with the red-colored QD until crosstalk correction.

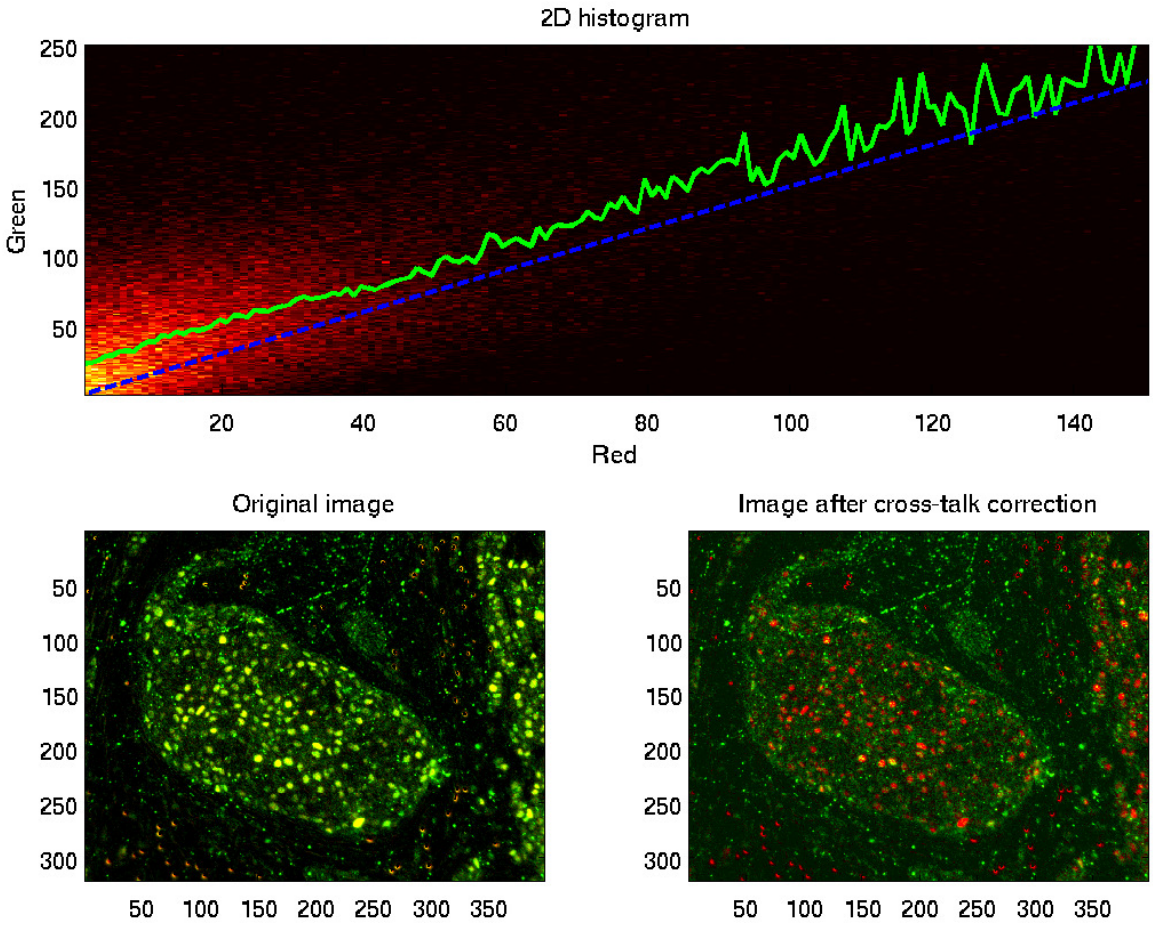

Figure 37: Observed Signal Crosstalk in Multiplexed Stained Tissues.
Figures show a 2-D histogram of normalized and unmixed QD signals. Lines show the estimated cross-talk among probes used. The first QD, shown in red, was applied before application of the second QD, shown in green.

Spectral Blurring and Chromatic Aberrations

Spectral blurring, also referred to as chromatic aberration, is a well-documented phenomenon whereby multicolored images appear in different focal planes according to the wavelength-dependent refraction (dispersion) of a lens. In its familiar form, chromatic aberration results in ghosted colored edges in photographs, or the appearance of rainbow colored stars in telescope images. For multiplexed QD multispectral images, this effect can be devastating because it interferes with precise spectral unmixing. Figure 39 is an excerpt of an unmixed and then pseudocolored photomicrograph of an RCC sample displaying two autofluorescence signatures in green or red. Chromatic aberration is visible in the image in the form of extra green halos surrounding the red objects (erythrocytes).

The broad acquisition bandwidth of multispectral QD data makes it especially susceptible to chromatic aberration and poorly suited to traditional hardware solutions to chromatic aberration which couple two dispersive media together. Software solutions for correcting aberration in RGB images, such as that proposed by Kaufmann et al. [97], exist and are included in most modern high-end consumer digital cameras. Such methods work by post-imaging modification (registration) of the red and blue images to match the green image. Extending this methodology to a multispectral image is also possible, although 30 registrations will be necessary to enhance the typical QD image stack.

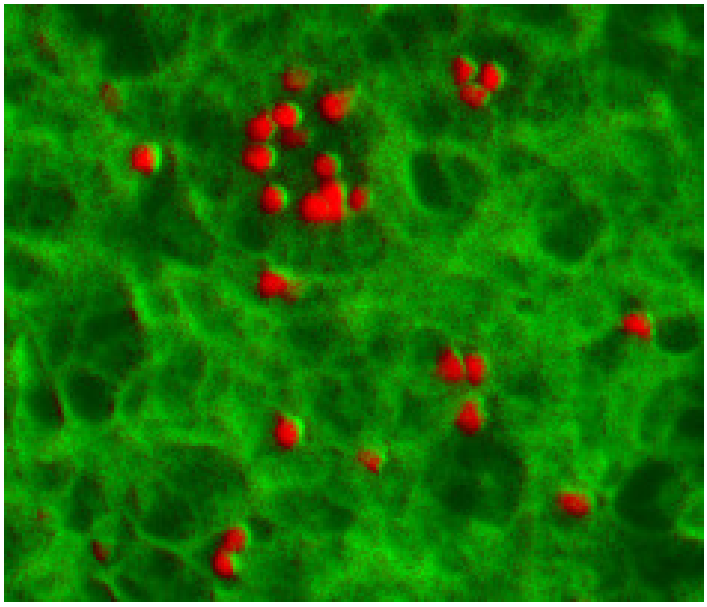

Figure 38: Pseudocolored Image Excerpt of RCC Tissue Autofluorescence which Demonstrates Chromatic Aberration.
Red blood cell autofluorescent signature is pseudocolored red, while renal tissue autofluorescence is pseudocolored green.

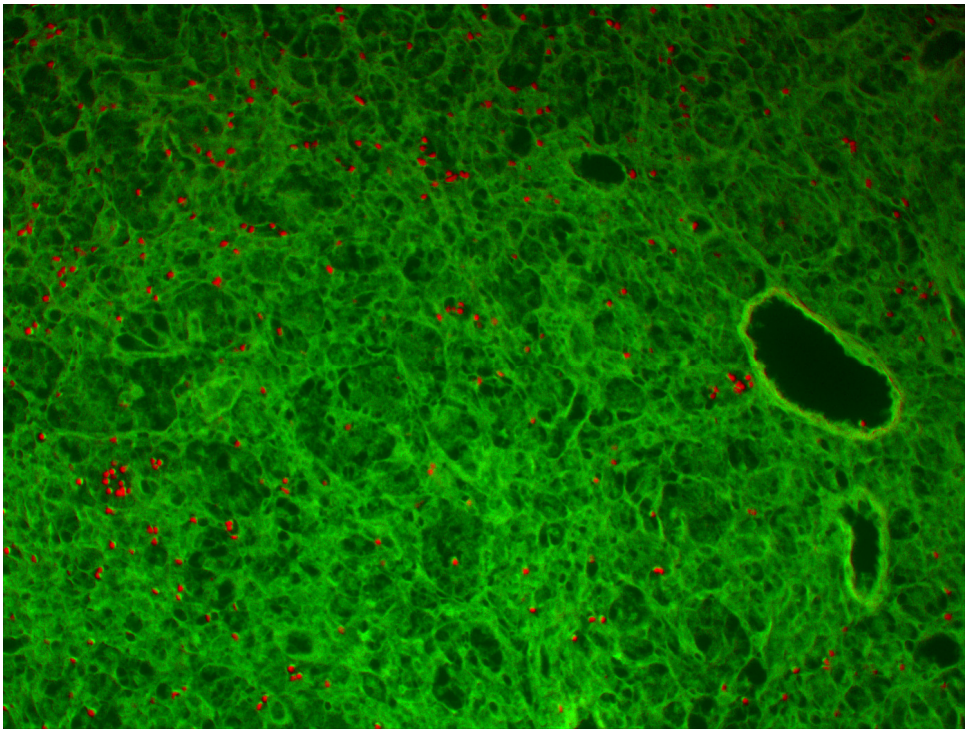

Figure 39: Pseudocolored Image of RCC Tissue Autofluorescence which Demonstrates Chromatic Aberration.
Red blood cell autofluorescent signature is pseudocolored red, while renal tissue autofluorescence is pseudocolored green.

As a precursor to correcting for aberration, a function describing the aberration must first be estimated. We begin by considering the image shown in Figure 39. A local correlation search was performed by selecting a window in the blue end of the spectrum and translating it locally many times. Each time we noted the correlation of the windowed blue spectrum image with the corresponding red image, which is held still. For each window, the direction and distance of translation which corresponds to the maximum correlation of overlapped windows was recorded. Because the image in question is composed largely of a broad-spectrum fluorescence, it is reasonable to expect the blue spectrum image to correlate well with the red spectrum image. This would not be the case for a multiplexed quantum dot image, for example. A color-coded map of the direction and magnitude of the estimated "maximum correlation" translations are shown in Figure 40.

As shown in Figure 40, the apparent chromatic blur was in a radial direction from the center of the image, with a magnitude that increases with distance from the center. This analysis should be repeated for each of the 31 spectral images, with one acting as a point of reference. After acquiring this empirical measure of the microscope's chromatic aberration for each wavelength, a geometric model of this aberration could be constructed. An example of such a function would be a linear increase with distance from the center, but more complicated models are possible, depending on the microscope setup, and focal plane. Once a suitable model is created for each wavelength image, they can each be modified according to the model in order to create a spectrally registered image stack.

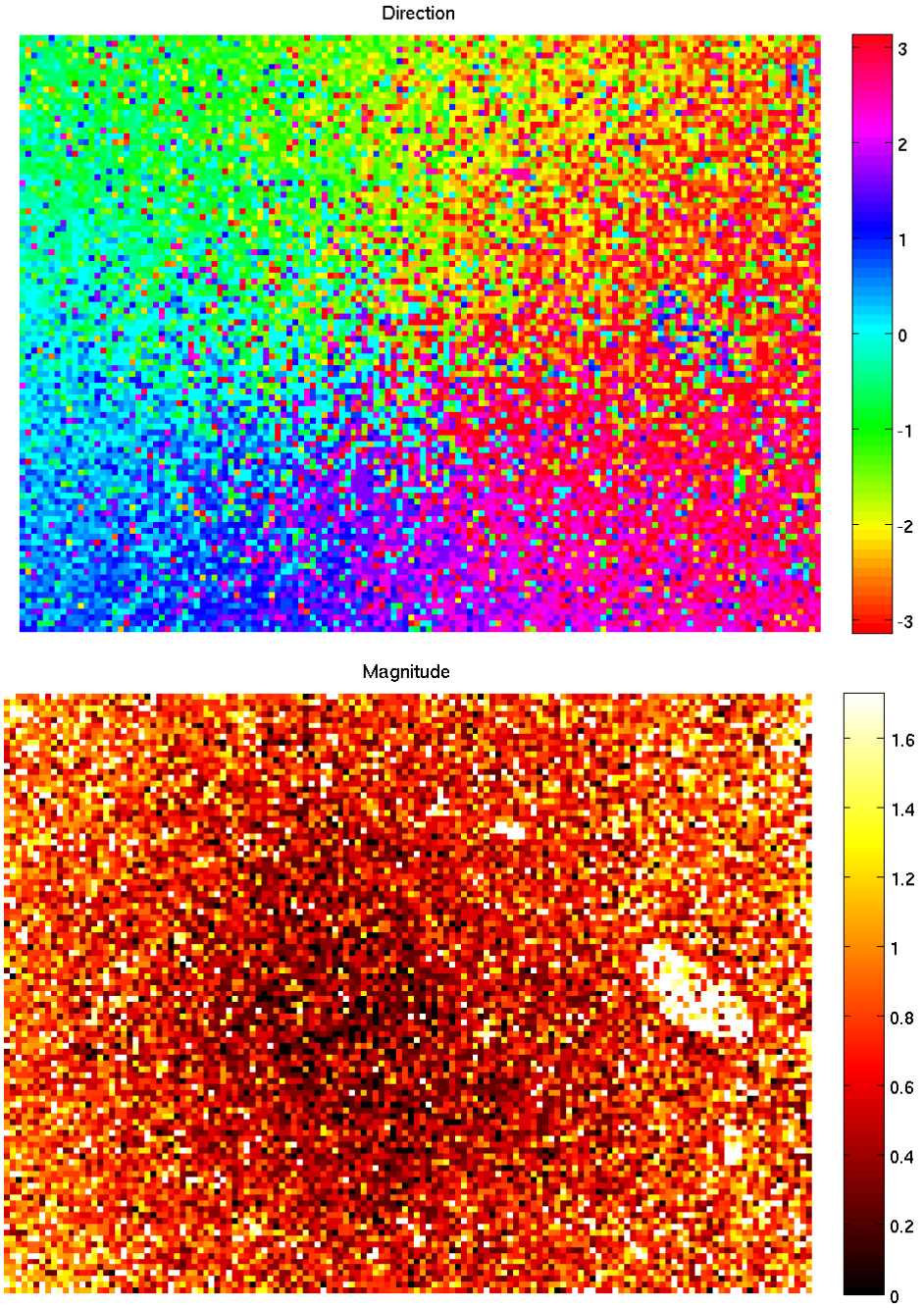

Figure 40: Map of Direction and Magnitude of Spectral Blurring.
Direction is given in radians, and distance in pixels.

Summary

In this chapter, we have discussed issues which currently prevent a reproducible and reliable quantitative multiplexed QD-IHC protocol from becoming a clinical reality. Despite current shortcomings, semi-quantitative results, such as those discussed in chapter 4, are currently achievable with proper controls and calibrations. Chief among outstanding issues is the physical aspect of the assay, which is having the intended reporter molecules attach to their intended targets in a linear and dose-dependent manner. In this respect, the mathematical and software solutions for proper unmixing and quantification appear to be more mature than the physical and biochemical aspects of the assay. For example, we have shown that characterization of the source spectra for these images is an achievable, even if sometimes overlooked, portion of protocol. Furthermore, we suggest that a simple model can be combined with current methods of spectral unmixing borrowed from other fields to achieve a reasonable result.

CHAPTER 6

CONCLUSIONS

To achieve the three specific aims of this research, three concrete deliverables were produced. These deliverables are: caCORRECT, validated biomarkers for RCC, and results of an investigation into the challenges QD-IHC. This chapter concludes the dissertation by offering a discussion of the current status of these deliverables, followed by a future outlook on these and related topics

Contributions to the Field

caCORRECT

In chapters 2 and 3, we have proposed and validated a microarray quality control system, caCORRECT. Descriptions of methods are found in this document, as well as in previous publication [11]. Much of the newer validation work is expected to be published concurrently in an appropriate peer-reviewed scientific journal. Source code for the version of caCORRECT described in this document is written entirely in MATLAB by me, with some recent help from Mitch Parry. Documentation of this code is maintained on our internal intralab wiki page, with the help of Mitch Parry. In addition to the MATLAB development version of caCORRECT, two other instances also exist. First, the Enterprise version of caCORRECT is implemented in PHP, JavaScript, C and C++ codes written mostly by myself and Todd Stokes., with help from JT Torrance and John Phan. The enterprise version exists as a web service, which has been freely available to the public at caCORRECT.bme.gatech.edu since the spring of 2007. The website is currently maintained by Sovandy Hang. The final implementation of caCORRECT is a grid service

as part of the National Cancer Institute's Cancer Bioinformatics Grid (caBIG). Conversion of caCORRECT into the caBIG grid service format and creation of associated silver-level compatibility review documents were done by Martin Ahrens and Todd Stokes.

Biomarkers

In chapter 4, Development of a Biomarker Based Diagnosis, we share the results of biomarker discovery, aided by caCORRECT, using a case study of Renal Cell Carcinoma (RCC) clinical samples. In this document, we focus on the thorough validation of two such markers, NNMT and PRKAB1, which allow 100% classification accuracy in 24 independent validation samples. In addition to these two well-studied markers, more than 10 other potential clinical markers have also been discovered during this process. These other markers represent the results of large scale collaborative study which are not appropriate for direct inclusion in this dissertation. Instead, these markers may be found in two other publications that I have contributed to: Phan et al, 2009 [98] or Osunkoya et al., 2009 [13].

Quantum Dot Methodology

As discussed in Chapter 5, QD-IHC has major potential as a clinical tool for sensitive and specific molecular diagnosis and prognosis of disease. Motivated by this potential, this dissertation has outlined key areas of improvement which must be made before translating QD-IHC to the clinic. Such a critical discussion is rare in the literature, which tends to focus on success rather than perceived failures in protocol. Despite these issues, data from existing duplex QD-IHC protocols have been used as a successful

demonstration of semi-quantitative comparison which can help separate two subtypes of RCC with moderate accuracy.

Future Outlook

caCORRECT

The current status of caCORRECT is a complete and functional system which includes many novel contributions. Further improvements of the various modules of caCORRECT are expected to be incremental in nature and have diminishing returns on investment. Such incremental impotents could be made by: (1) increasing the complexity of artifact masking procedures, using an array of kernels designed for common artifacts, (2) incorporating the complexity of existing 3rd party gene expression models into caCORRECT's modeling procedure, or (3) further investigate convergence of multiple rounds of artifact detection, and testing the possibility of fuzzy artifact classification. Originally, these ideas were shelved to accommodate the user's need for speed, and these issues have yet to subside as the size of microarray datasets seems to be scaling along with computer hardware capabilities.

Aside from these incremental improvements, the broader impact of caCORRECT could be improved by porting the same methodologies to other array platforms or technologies. We have already accomplished a proof of concept for the Illumina BeadChip platform [10], but many other applications exist, including the emerging field of QC for next-generation sequencing technologies. The pending incorporation of caCORRECT as a caBIG grid service is expected to help increase broader impact as well.

To increase the impact and use of caCORRECT in the community, we must show that caCORRECT has a retroactive impact on the quality of previous microarray experiments. To show this, we will now revisit the results of previously high-impact

published papers to see if good biomarkers can be found after caCORRECT treatment of raw data.

On another frontier, Todd Stokes' system, ArrayWiki [57], is now the premiere platform for showcasing caCORRECT, and it is also responsible for the most uses of caCORRECT, beyond even my own use. ArrayWiki is an open wiki repository for microarray data which includes mandatory preprocessing with caCORRECT for all new data imports. Importantly, ArrayWiki also displays heat maps from caCORRECT for every microarray. As ArrayWiki grows to incorporate a higher percentage of the world's microarray data, so grows the impact of caCORRECT.

Biomarkers

Here, we have highlighted the discovery of a two-gene panel of biomarkers for subtyping between two classes of RCC. Clearly, much work is left to be done before we are able to claim a truly clinically relevant test for RCC. Specifically, we must develop a panel of biomarkers that is suitable for the differentiation of all five of the common subtypes of RCC, and not just CC and CHR. From the available PCR data which was collected as a byproduct of proving caCORRECT's effect on reliability of biomarker selection, we are able to create such a panel, but it is yet to be verified as reliable based on independent clinical samples. Future work will involve creating optimal sets of biomarkers for simultaneous differentiation of multiple clinical subtypes of RCC.

Although such a PCR-based test would be valuable indeed, the search for more sensitive techniques of biomarker-based analysis is also ongoing. This includes not only the QD-IHC discussed here, but also bimolecularly specific methods of circulating tumor cell [74, 99-104] or blood-bourn biomarker detection [105, 106]. Notable failures of reproducibility in these fields highlight the need for QC in all aspects of biomarker identification [107-110].

Quantum Dots as a Clinical Technology

The reliable quantification of QD-IHC is of critical importance to its clinical translation. To achieve this, we must first start with a basic assessment of QD quantification in a cell and antibody free environment. Some work has already been done to assess the dynamic range and detection sensitivity limit of QDs in large beads [111], but not necessarily in the single antibody-conjugated nanoparticle form that is the most likely form for a clinical application.

Our own observations, as well as those of others [112], have confirmed that QDs can undergo time-dependent photo brightening or photo darkening. Such time-dependent effects represent a significant hurdle to quantification, and thus work to assess and stabilize QD signal over time are needed. Similarly, there is also a need to increase the relative brightness of QD signal with respect to tissue autofluorescence [74, 76]. Some progress has been made to increase signal to background using the unique properties of QD i.e. long emission half-life [79], or relative photostabillity [75, 80]. Unfortunately, most of the existing solutions for increasing the QD signal to autofluorescence signal ratio involve techniques which also affect the brightness of QDs, and are thus difficult to apply directly to a quantitative protocol.

Aside from these issues related to the optical properties of QDs, quantitative IHC even without QDs has its own problems. Although some systems, such as AQUA [113] have been proposed to quantify staining, reproducibility outside of highly-controlled tissue microarrays has yet to be demonstrated. Using improper antibody concentrations which cause nonlinear response to antigen has been linked to failure in a past study, notably by the developers of AQUA [114]. To combat these common issues of IHC and QD-IHC, we must conduct basic sample-free antigen capture experiments to verify and quantify linear binging of probes to targets.

The capacity of QD-IHC to be highly multiplexed invites further room for quality control. In addition to limiting antibody cross-reactivity, more basic issues of QD-to-

target binding persistence must be investigated with respect to the increased number of washing and incubation steps required in multiplexed assays. To do this, we have begun to investigate sample-free serial binding and disassociation kinetics.

Quantum dots are not the only emerging technology which shows promise for multiplexed analyses of tissue biopsies; Surface Enhanced Raman Spectroscopy (SERS) nanotags have also been proposed for in-situ protein analysis of tissue specimens [115-117]. Due to their complex spectra, SERS tags have a higher potential for molecular specificity during imaging, but also require more sensitive instrumentation, and more computation to achieve these advantages. Just like QD tags, SERS tags face the same problems of antibody cross talk, yet they have the potential to be more easily separated from background tissue autofluorescence signals. SERS work should be able to borrow many spectral unmixing solutions from another spectral tissue imaging technology, Imaging Mass Spectrometry (IMS) [3, 118-122]. IMS has the advantage of being antibody-free, but is severely limited by spatial resolution as well as high cost—making it currently unsuitable for clinical application.

Closing Remarks

Both the development and implementation of a clinical biomarker test require quality control to ensure efficiency and reliability. Examples of this include insurance of the best possible microarray data being fed into the pipeline, as well as reproducible quantification of IHC results at the point of clinical contact. This dissertation represents a significant step forward in achieving these goals of quality, reliability, and reproducibility of each piece of the translational bioinformatics pipeline. Future work is expected to build on these advances to deliver the best possible quality of diagnosis to patients and their physicians.

APPENDIX A

SOLUTION OF GENE EXPRESSION MODEL

The model which describes the relationship between observed probe intensities (raw chip data) and gene expression (data for modeling) is described by the following equation, previously given in chapter 2.

$$x_{b,p,j} = \theta_{p,j} a_{b,p} \tag{A.1}$$

In the above model equation, $x_{b,p,j}$ is the observed intensity for the b^{th} probe in the p^{th} probe set on the j^{th} chip, $\theta_{p,j}$ is the gene expression term corresponding to target RNA concentration, and $a_{b,p}$ is the lumped probe affinity term. This appendix discusses the ways in which this equation is practically solved.

Discussion of Error Term

This gene expression model can be interpreted with different error models. In each case, $\varepsilon_{b,p,j}$ and or $\varepsilon'_{b,p,j}$ will represent the error terms which contribute to the observed intensity $x_{b,p,j}$. One error model option is additive in the domain of probe intensity. In this case, $\varepsilon_{b,p,j}$ represents noise that is independent of the intensity signal.

$$x_{b,p,j} = \theta_{p,j} a_{b,p} + \varepsilon_{b,p,j} \tag{A.2}$$

Another interpretation is that error is multiplicative in the domain of probe intensity, which is equivalent to additive in the log domain of probe intensity. In this case, $\varepsilon'_{b,p,j}$ represents noise that is proportionate to the intensity signal.

$$x_{b,p,j} = \theta_{p,j} a_{b,p} \varepsilon'_{b,p,j} \tag{A.3}$$

-or-

$$\log(x_{b,p,j}) = \log(\theta_{p,j}) + \log(a_{b,p}) + \log(\varepsilon'_{b,p,j}) \tag{A.4}$$

In reality, the noise may be best modeled as a combination of these two types of additive and multiplicative noise.

$$x_{b,p,j} = \theta_{p,j} a_{b,p} \varepsilon'_{b,p,j} + \varepsilon_{b,p,j} \tag{A.5}$$

Figure 41 shows an example of real data from the Young dataset for a single probe set selected for its relatively large dynamic range. Figure 42, Figure 43, and Figure 44 show simulated probe data in the same range as Figure 41 with additive, multiplicative and combination models of noise, respectively. caCORRECT uses an additive model by default, but options exist to use a multiplicative model as well. The additive noise model is chosen over the multiplicative noise model for superior stability in the face of low-expressing genes.

According to the model, data from each probe set are independent of one another, and contain no common terms. Thus, reconstructing the probe intensity model equation for each observed probe intensity on each chip in a dataset produces a series of P independent sets of equations, where P is the total number of probe sets represented on the microarray platform. How these sets of equations are solved depends on the error model used.

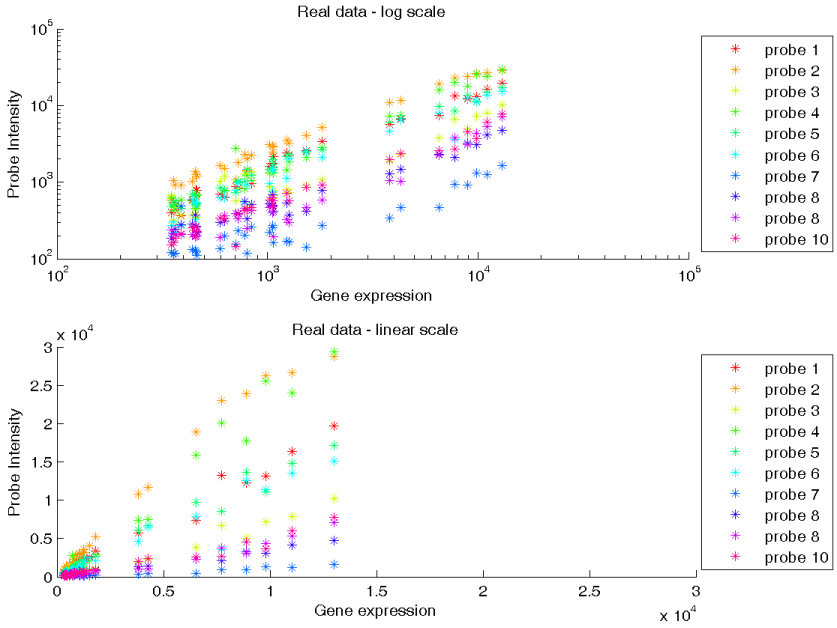

Figure 41: Plot of Gene Expression Versus Probe Intensity for Real Data.
A mixture of intensity dependent and independent noise is observable.

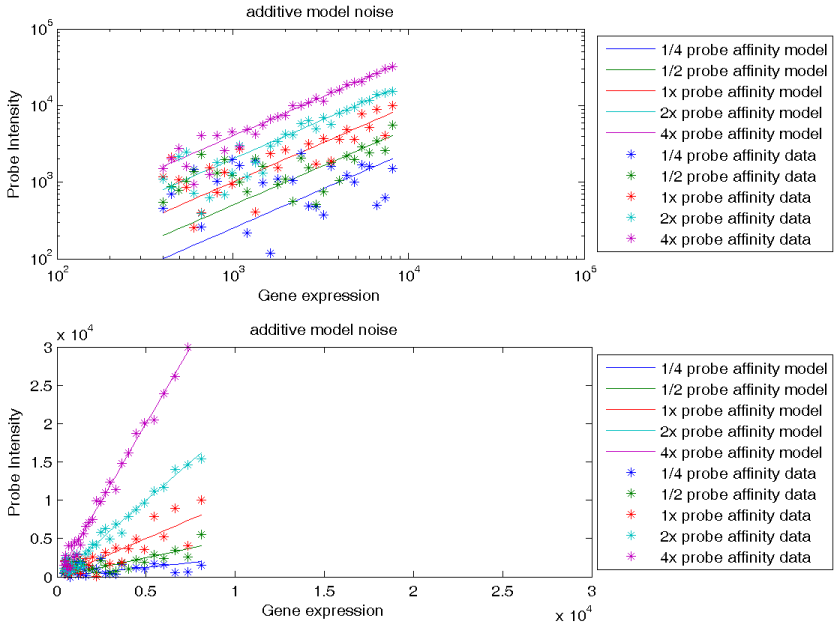

Figure 42: Plot of Gene Expression Versus Probe Intensity for Simulated Data with Additive Noise.
Residuals converge for high intensities when viewed on the logarithmic axes.

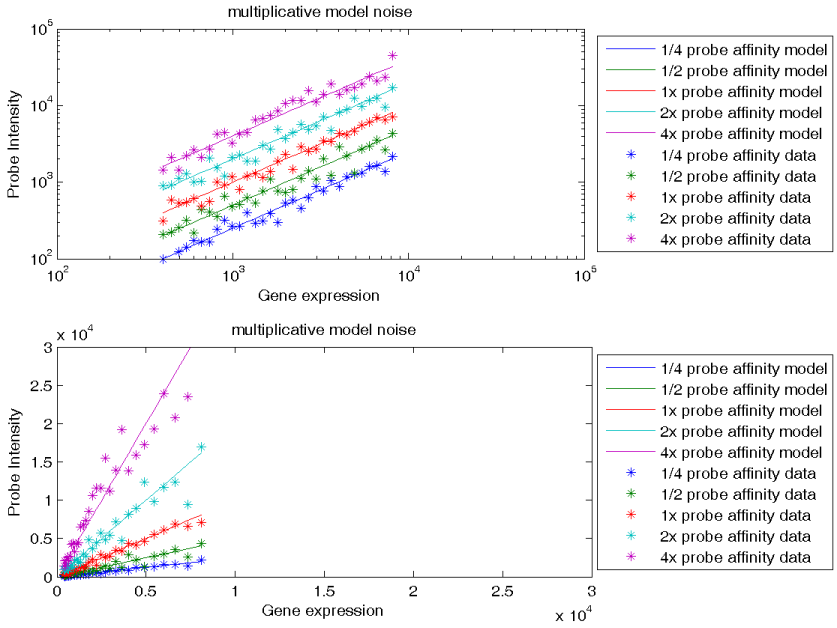

Figure 43: Plot of Gene Expression Versus Probe Intensity for Simulated Data with Multiplicative Noise.
Residuals converge for low intensities when viewed on the linear axes.

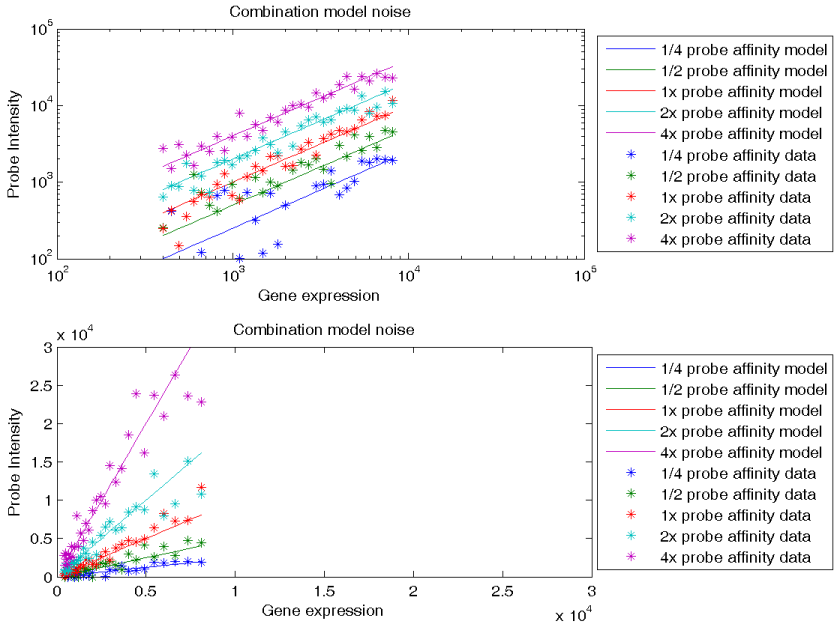

Figure 44: Plot of Gene Expression Versus Probe Intensity for Simulated Data with a Mix of Additive and Multiplicative Noise.
Residuals do not appear to converge for either of the scaled axes.

Solution of Additive Gene Expression Model

The set of equations for the p^{th} probe set using an additive model of error can be represented in the following matrix form, given a set of N chips and B_p probes in the p^{th} probe set.

$$\begin{bmatrix} x_{1,p,1} & \cdots & x_{B_p,p,1} \\ \vdots & \ddots & \vdots \\ x_{1,p,N} & \cdots & x_{B_p,p,N} \end{bmatrix} = \begin{bmatrix} \theta_{p,1} \\ \vdots \\ \theta_{p,N} \end{bmatrix} \begin{bmatrix} a_{1,p} & \cdots & a_{B_p,p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,p,1} & \cdots & \varepsilon_{B_p,p,1} \\ \vdots & \ddots & \vdots \\ \varepsilon_{1,p,N} & \cdots & \varepsilon_{B_p,p,N} \end{bmatrix} \quad \text{(A.7)}$$

For ease of notation, we will refer to this equation in a condensed form as follows.

$$\mathbf{X}_p = \boldsymbol{\theta}_p \mathbf{a}_p + \boldsymbol{\varepsilon}_p \quad \text{(A.8)}$$

We define the solution to this matrix equation as that which minimizes the Frobenius norm of the above error matrix $\boldsymbol{\varepsilon}_p$ as defined below.

$$\left\| \boldsymbol{\varepsilon}_p \right\|_F = \sqrt{\sum_{j=1}^{N} \sum_{b=1}^{B_p} \left(\varepsilon_{b,p,j} \right)^2} \quad \text{(A.9)}$$

To be able to come up with a unique solution, we introduce the constraint that the geometric mean of the lumped probe affinity terms $a_{b,p}$ equals one. The number one is arbitrary here, but it allows the convenient interpretation that the values of gene expression, $\theta_{p,j}$, are on the same scale as the probe intensities, $x_{b,p,j}$.

$$1 = \sqrt[B_p]{\left(\prod_{b=1}^{B_p} a_{b,p} \right)} \quad \text{(A.10)}$$

The solution which satisfies the above conditions can be derived from the singular value decomposition (SVD) of $\mathbf{X}_p$. Here, the SVD of $\mathbf{X}_p$ is given in the form of $\mathbf{X}_p = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$, such that $\mathbf{U} \in \Re^{N \mathrm{x} N}$, $\mathbf{S} \in \Re^{N \mathrm{x} B_p}$, and $\mathbf{V} \in \Re^{B_p \mathrm{x} B_p}$. If the largest singular value in **S**, s_1, is arranged as the first diagonal element of **S**, then $\boldsymbol{\theta}_p$ is s_1 times the first column of **U** and $\mathbf{a}_p$ is the first column of **V**. Once derived from the SVD, $\mathbf{a}_p$ and $\boldsymbol{\theta}_p$ may then

be scaled by respective multiplication and division by a single factor in order to satisfy the geometric mean constraint as stated earlier.

If we assume that $\mathbf{X}_p$ is rank 1, (that it can be perfectly approximated by the multiplication of a single column vector $\boldsymbol{\theta}_p$ and single row vector $\mathbf{a}_p$), then we can use a shortcut to estimate $\boldsymbol{\theta}_p$ and $\mathbf{a}_p$ directly from $\mathbf{X}_p$ without using an SVD. In such a situation the summation of the b^{th} column of $\mathbf{X}_p$ would give a clue to the b^{th} element of $\mathbf{a}_p$ as follows.

$$\sum_{j=1}^{N}\left(x_{b,p,j}\right)=\sum_{j=1}^{N}\left(\theta_{p,j}a_{b,p}\right)=a_{b,p}\sum_{j=1}^{N}\left(\theta_{p,j}\right) \tag{A.11}$$

This b^{th} column summation of $\mathbf{X}_p$ is defined as $\varphi_{b,p}$, which can then be combined into a row vector, $\boldsymbol{\varphi}_p$.

$$\boldsymbol{\varphi}_p=\begin{bmatrix}\varphi_{1,p} & \cdots & \varphi_{B_p,p}\end{bmatrix}=\begin{bmatrix}a_{1,p} & \cdots & a_{B_p,p}\end{bmatrix}\sum_{j=1}^{N}\left(\theta_{p,j}\right) \tag{A.12}$$

In this form, we can see that $\boldsymbol{\varphi}_p$ is simply a scaled version of $\mathbf{a}_p$. Because of the geometric mean constraint on $\mathbf{a}_p$, $\mathbf{a}_p$ can be easily derived from $\boldsymbol{\varphi}_p$ by normalizing $\boldsymbol{\varphi}_p$ by its geometric mean as follows.

$$\mathbf{a}_p=\frac{\boldsymbol{\varphi}_p}{\sqrt[Bp]{\prod_{b=1}^{B_p}\varphi_{b,p}}} \tag{A.13}$$

Similarly, summation of the j^{th} row of $\mathbf{X}_p$ would give a clue to $\theta_{p,j}$.

$$\sum_{b=1}^{B_p}\left(x_{b,p,j}\right)=\sum_{b=1}^{B_p}\left(\theta_{p,j}a_{b,p}\right)=\theta_{p,j}\sum_{b=1}^{B_p}\left(a_{b,p}\right) \tag{A.14}$$

Since the summation of $\mathbf{a}_p$ is calculable, so is each element of $\boldsymbol{\theta}_p$.

$$\theta_{p,j} = \frac{\sum_{b=1}^{B_p} \left(x_{b,p,j} \right)}{\sum_{b=1}^{B_p} \left(a_{b,p} \right)} \tag{A.15}$$

Incorporation of artifacts into this model is done in the following 2-step iterative procedure similar to the Expectation-Maximization algorithm. This procedure, given below, is repeated until $\boldsymbol{\theta}_p$ and $\mathbf{a}_p$ converge

1) Estimate $\boldsymbol{\theta}_p$ and $\mathbf{a}_p$ given $\mathbf{X}_p$, either using the SVD or the mean approximation as described previously.
2) Replace known artifact values in $\mathbf{x}_p$ with information from the corresponding elements of $\boldsymbol{\theta}_p \mathbf{a}_p$.

Solution of Multiplicative Gene Expression Model

The set of equations for the p^{th} probe set using a multiplicative model of error can be represented in the following matrix form:

$$\log\left(\begin{bmatrix} \begin{bmatrix} x_{1,p,1} \\ \vdots \\ x_{1,p,N} \end{bmatrix} \\ \begin{bmatrix} x_{2,p,1} \\ \vdots \\ x_{2,p,N} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x_{B_p,p,1} \\ \vdots \\ x_{B_p,p,N} \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix} \\ \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} & \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \end{bmatrix} \\ \vdots & \vdots \\ \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} & \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} \end{bmatrix} \log\left(\begin{bmatrix} \theta_{p,1} \\ \vdots \\ \theta_{p,N} \\ a_{1,p} \\ \vdots \\ a_{B_p,p} \end{bmatrix}\right) + \log\left(\begin{bmatrix} \begin{bmatrix} \varepsilon'_{1,p,1} \\ \vdots \\ \varepsilon'_{1,p,N} \end{bmatrix} \\ \begin{bmatrix} \varepsilon'_{2,p,1} \\ \vdots \\ \varepsilon'_{2,p,N} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \varepsilon'_{B_p,p,1} \\ \vdots \\ \varepsilon'_{B_p,p,N} \end{bmatrix} \end{bmatrix}\right)$$

(A.16)

For ease of notation, we will refer to this equation in a condensed form as follows.

$$\mathbf{x}'_p = \mathbf{M}_p \begin{bmatrix} \boldsymbol{\theta}'_p \\ \cdots \\ \mathbf{a}'^{\mathrm{T}}_p \end{bmatrix} + \boldsymbol{\varepsilon}'_p \tag{A.17}$$

Like in the additive noise model, we must also introduce a constraint on $\mathbf{a}'_p$ to achieve a unique solution. In the case of multiplicative noise, however, we represent this same constraint in a more convenient form.

$$0 = \frac{1}{B_p} \sum_{b=1}^{B_p} \log(a_{b,p}) \tag{A.18}$$

In this form, the constraint can easily be incorporated into the existing matrix equation structure by appending a zero to the bottom of $\mathbf{x}'_p$ and a row consisting of N zeros, followed by B_p ones to the bottom of $\mathbf{M}_p$. In this augmented form, it is straightforward to estimate $\begin{bmatrix} \boldsymbol{\theta}'_p \\ \cdots \\ \mathbf{a}'^{\mathrm{T}}_p \end{bmatrix}$ given data from $\mathbf{x}'_p$, and the fixed coefficient matrix, $\mathbf{M}_p$ in a way that minimizes an error which is the Frobenius norm of $\boldsymbol{\varepsilon}'_p$. The Frobenius norm of $\boldsymbol{\varepsilon}'_p$ is defined in the following equation.

$$\left\| \boldsymbol{\varepsilon}'_p \right\|_F = \sqrt{\sum_{j=1}^{N} \sum_{b=1}^{B_p} \left(\log(\varepsilon'_{b,p,j}) \right)^2} \tag{A.19}$$

The solution to this error minimization is given by the following equation, which uses the pseudo-inverse of $\mathbf{M}_\mathrm{p}$. The superscript $^{\mathrm{T}}$ represents the matrix transpose.

$$\left(\mathbf{M}_p{}^{\mathrm{T}} \mathbf{M}_p \right)^{-1} \mathbf{M}_p{}^{\mathrm{T}} \mathbf{x}'_p = \begin{bmatrix} \boldsymbol{\theta}'_p \\ \cdots \\ \mathbf{a}'^{\mathrm{T}}_p \end{bmatrix} \tag{A.20}$$

Note that because the constraint on $\mathbf{a}'_p$ is completely achievable, we do not need to worry about it being codified here as a "soft" constraint as part of a least squares optimization.

Incorporation of known artifacts into this model is also rather straightforward. One way to ignore artifacts is to simply excise the rows of $\mathbf{x}'_p$ and $\mathbf{M}_p$ which contain

artifactual data. In rare cases, however, noisy data can lead to excessive excision that can in turn result in $\mathbf{M}_p{}^{\mathrm{T}}\mathbf{M}_p$ being singular, and so this method is not preferred. An alternative method is to multiply the rows of $\mathbf{x}'_p$ and $\mathbf{M}_p$ which contain artifactual data by a small number, such as the 0.1 used by caCORRECT. This has the effect of down-weighting any residuals derived from artifact data during the optimization. Because the data are not completely excised, $\mathbf{M}_p{}^{\mathrm{T}}\mathbf{M}_p$ remains invertible even in the case of extensive artifact coverage.

REFERENCES

1. Shi, L., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **24**(9): p. 1151-61.

2. Young, A.N., et al., *Expression Profiling of Renal Epithelial Neoplasms A Method for Tumor Classification and Discovery of Diagnostic Molecular Markers*. 2001, ASIP.

3. Muir, E.R., et al., *Multivariate Analysis of Imaging Mass Spectrometry Data.* Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on, 2007: p. 472-479.

4. Affymetrix, *Statistical Algorithms Description Document* 2002.

5. Irizarry, R.A., et al., *Summaries of affymetrix GeneChip probe level data.* Nucleic Acids Research, 2003. **31**(4): p. -.

6. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**: p. 249-264.

7. Affymetrix, *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.* 2005.

8. Moffitt, R., et al., *Effect of Outlier Removal on Gene Marker Selection Using Support Vector Machines.* Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, 2005: p. 917-920.

9. Moffitt, R.A., et al., *Simple Outlier Removal Improves the Performance of Support Vector Machines as a Biomarker Selection Method.* Intelligent Systems for Molecular Biology Conference 2005, Detroit, MI, 2005.

10. Stokes, T.H., et al., *Extending microarray quality control and analysis algorithms to Illumina chip platform.* Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007: p. 4637-4640.

11. Stokes, T.H., et al., *chip artifact CORRECTion (caCORRECT): A Bioinformatics System for Quality Assurance of Genomics and Proteomics Array Data.* Annals of Biomedical Engineering, 2007. **35**(6): p. 1068-1080.

12. Torrance, J.H., et al., *Can We Trust Biomarkers? Visualization and Quantification of Outlier Probes in High Density Oligonucleotide Microarrays.* Life Science Systems and Applications Workshop, 2007. IEEE/NIH BISTI, 2007: p. 251-254.

13. Osunkoya, A., et al., *Diagnostic biomarkers for renal cell carcinoma: selection using novel bioinformatics systems for microarray data analysis.* Human Pathology (In Press), 2009.

14. Yin-Goen, Q., et al., *Advances in molecular classification of renal neoplasms.* Histology and histopathology, 2006. **21**(1-3): p. 325-339.

15. Moffitt, R.A., et al., *Quality Control of Highly Multiplexed Proteomic Immunostaining with Quantum Dots: Correcting for Crosstalk.* Engineering in Medicine and Biology Society, 2009. IEEE-EMBS 2009. 31st Annual International Conference of the, 2009.

16. Caldwell, M.L., et al., *Simple quantification of multiplexed Quantum Dot staining in clinical tissue samples.* Engineering in Medicine and Biology Society, 2008. 30th Annual International Conference of the IEEE, 2008: p. 1907-1910.

17. Moffitt, R. and M. Wang, *Microarray Data Analysis.* Wiley Encyclopedia of Biomedical Engineering, 2006. **4**: p. 2270-2279.

18. Ancona, N., et al., *On the statistical assessment of classifiers using DNA microarray data.* BMC Bioinformatics, 2006. **7**: p. 387.

19. Baildi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-424.

20. Brettschneider, J., et al., *Quality Assessment for Short Oligonucleotide Microarray Data. Rejoinder.* Technometrics, 2008. **50**(3): p. 279.

21. Cortes, C. and M. Mohri. *AUC optimization vs error rate minimization*. in *Advances in Neural Information Processing Systems (NIPS)*. 2003.

22. Davis, C.A., et al., *Reliable gene signatures for microarray classification: assessment of stability and performance.* Bioinformatics, 2006. **22**(19): p. 2356.

23. Dupuy, A. and R.M. Simon, *Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting.* JNCI Journal of the National Cancer Institute, 2007. **99**(2): p. 147.

24. Ferri, C., J. Hernandez-Orallo, and R. Modroiu, *An experimental comparison of performance measures for classification.* Pattern Recognition Letters, 2009. **30**: p. 27-38.

25. Hoffmann, R., T. Seidl, and M. Dugas, *Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis.* Genome Biol, 2002. **3**(7): p. RESEARCH0033.

26. Ioannidis, J.P., et al., *Repeatability of published microarray gene expression analyses.* Nat Genet, 2009. **41**(2): p. 149-55.

27. Jenssen, T.K., et al., *Analysis of repeatability in spotted cDNA microarrays.* Nucleic Acids Research, 2002. **30**(14): p. 3235.

28. Lee, M.L.T., et al., *Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(18): p. 9834-9839.

29. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* Lancet, 2005. **365**(488-492).

30. Miller, L.D., et al., *Optimal gene expression analysis by microarrays.* Cancer Cell, 2002. **2**(5): p. 353-361.

31. Ntzani, E. and J. Ioannidis, *Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.* Lancet, 2003. **362**(9394): p. 1439-1444.

32. Quackenbush, J., *Microarray Analysis and Tumor Classification.* NEJM, 2006. **354**(23): p. 2463-2472.

33. Shedden, K., et al., *Gene expression\[ndash]\ based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nature Medicine, 2008. **14**(8): p. 822-827.

34. Simon, R., et al., *Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification.* Journal of the National Cancer Institute, 2003. **95**(1): p. 14-18.

35. Varma, S. and R. Simon, *Bias in error estimation when using cross-validation for model selection.* BMC Bioinformatics, 2006. **7**(91).

36. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nature Methods, 2005. **2**(5): p. 345-349.

37. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science(Washington), 1995. **270**(5235): p. 467-470.

38. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME) - toward standards for microarray data.* Nature Genetics, 2001. **29**(4): p. 365-371.

39. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.* Proceedings of the National Academy of Sciences, 2001. **98**(1): p. 31.

40. Reimers, M. and J.N. Weinstein, *Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases.* Bmc Bioinformatics, 2005. **6**: p. -.

41. Suarez-Farinas, M., A. Haider, and K.M. Wittkowski, *"Harshlighting" small blemishes on microarrays.* Bmc Bioinformatics, 2005. **6**: p. -.

42. Sasik, R., E. Calvo, and J. Corbeil, *Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model.* Bioinformatics, 2002. **18**(12): p. 1633-1640.

43. Sidorov, I.A., et al., *Oligonucleotide microarray data distribution and normalization.* Information Sciences, 2002. **146**(1-4): p. 67-73.

44. Yang, M.C.K., et al., *A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays.* Physiological Genomics, 2001. **7**(1): p. 45-53.

45. Lu, C., *Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays.* Bmc Bioinformatics, 2004. **5**: p. -.

46. Hill, A.A., et al., *Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls.* Genome Biol, 2001. **2**(12): p. RESEARCH0055.

47. Henning, P.A., et al., *ChipQC: Microarray Artifact Visualization Tool*, in *Intelligent Systems for Molecular Biology*. 2005: Detroit, MI.

48. Buness, A., et al., *arrayMagic: two-colour cDNA microarray quality control and preprocessing*. 2005, Oxford Univ Press. p. 554-556.

49. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.

50. Suárez-Fariñas, M., et al., *Harshlight: a" corrective make-up" program for microarray chips.* BMC Bioinformatics, 2005. **6**(1): p. 294.

51. Fink, G.R., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Molecular biology of the cell, 1998. **9**(12): p. 3273-3297.

52. Kerr, M.K. and G.A. Churchill, *Experimental design for gene expression microarrays.* Biostatistics, 2001. **2**(2): p. 183-201.

53. Brodsky, L., et al., *Identification and handling of artifactual gene expression profiles emerging in microarray hybridization experiments.* Nucleic Acids Research, 2004. **32**(4): p. -.

54. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nature Biotechnology, 2003. **21**(7): p. 818-821.

55. Wu, Z., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.

56. Kluger, *Relationship between gene co-expression and probe localization on microarray slides.* BMC Genomics, 2003.

57. Stokes, T., et al., *ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses.* BMC Bioinformatics, 2008. **9**(Suppl 6): p. S18.

58. Bolstad, B.M. *PLM Image Gallery*. 2009 [cited 2009 June 1]; Available from: plmimagegallery.bmbolstad.com.

59. Cope, L.M., et al., *A benchmark for Affymetrix GeneChip expression measures*. 2004, Oxford Univ Press. p. 323-331.

60. Li, C. and W.H. Wong, *DNA-chip analyzer (dChip).* The analysis of gene expression data: methods and software. New York: Springer, 2003. **504**.

61. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2001: Wiley New York.

62. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays*. 2008, Google Patents.

63. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proceedings of the National Academy of Sciences, 2002. **99**(10): p. 6567.

64. Cristianini, N. and J. Shawe-Taylor, *An introduction to Support Vector Machines.* 2000.

65. Jirapech-Umpai, T. and S. Aitken, *Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes.* feedback, 2005.

66. Phan, J.H., et al., *Improving Microarray Sample Size Using Bootstrap Data Combination.* Computer and Computational Sciences, 2008. IMSCCS'08. International Multisymposiums on, 2008: p. 37-44.

67. Efron, B., *Estimating the error rate of a prediction rule: improvement on cross-validation.* Journal of the American Statistical Association, 1983: p. 316-331.

68. Efron, B. and R. Tibshirani, *Improvements on cross-validation: the. 632+ bootstrap method.* Journal of the American Statistical Association, 1997: p. 548-560.

69. Ambroise, C. and G. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data.* PNAS, 2002. **99**(10): p. 6562-6566.

70. Slamon, D.J., et al., *Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2.* New England Journal of Medicine, 2001. **344**(11): p. 783.

71. Chan, W.C.W., et al., *Luminescent quantum dots for multiplexed biological detection and imaging.* Current Opinion in Biotechnology, 2002. **13**(1): p. 40-46.

72. Gao, X. and S. Nie, *Molecular profiling of single cells and tissue specimens with quantum dots.* Trends in Biotechnology, 2003. **21**(9): p. 371-373.

73. Han, M., et al., *Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules.* Nature Biotechnology, 2001. **19**: p. 631-635.

74. Smith, A.M., et al., *Multicolor quantum dots for molecular diagnostics of cancer.* Expert Rev. Mol. Diagn., 2006. **6**(2): p. 231-244.

75. Xing, Y., et al., *Bioconjugated quantum dots for multiplexed and quantitative immunohistochemistry.* Nature Protocols, 2007. **2**(5): p. 1152-1165.

76. Yezhelyev, M.V., et al., *Emerging use of nanoparticles in diagnosis and treatment of breast cancer.* Lancet Oncol, 2006. **7**(8): p. 657-667.

77. Goldman, E.R., et al., *Multiplexed toxin analysis using four colors of quantum dot fluororeagents.* Anal. Chem, 2004. **76**(3): p. 684-688.

78. Wu, X., et al., *Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots.* Nature Biotechnology, 2002. **21**(1): p. 41-46.

79. Dahan, M., et al., *Time-gated biological imaging by use of colloidal quantum dots.* Optics Letters, 2001. **26**(11): p. 825-827.

80. Sukhanova, A., et al., *Biocompatible fluorescent nanocrystals for immunolabeling of membrane proteins and cells.* Analytical biochemistry, 2004. **324**(1): p. 60-67.

81. Mansfield, J.R., et al., *Autofluorescence removal, multiplexing, and automated analysis methods for in-vivo fluorescence imaging.* Journal of Biomedical Optics, 2005. **10**: p. 041207.

82. Levenson, R.M., *Spectral imaging and pathology: seeing more.* Lab Medicine, 2004. **35**(4): p. 244.

83. Fare, T.L., et al., *Effects of atmospheric ozone on microarray data quality.* Analytical Chemistry, 2003. **75**(17): p. 4672-4675.

84. Bolstad, B., *Probe level quantile normalization of high density oligonucleotide array data.* Unpublished Manuscript, 2001.

85. Schuetz, A.N., et al., *Molecular Classification of Renal Tumors by Gene Expression Profiling*. 2005, ASIP. p. 206-218.

86. West, M., et al., *Predicting the clinical status of human breast cancer by using gene expression profiles.* Proceedings of the National Academy of Sciences, 2001. **98**(20): p. 11462.

87. Vajda, I., *Theory of Information and Statistical Decision.* Alfa, Bratislava, 1981.

88. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nature Medicine, 2002. **8**(1): p. 68-74.

89. Chuaqui, R.F., et al., *Post-analysis follow-up and validation of microarray experiments.* NATURE GENETICS, 2002. **32**(supp): p. 509-514.

90. Simon, R., *Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data.* Br J Cancer, 2003. **89**: p. 1599-1604.

91. Phan, J., et al., *Improvement of SVM Algorithm for Microarray Analysis Using Intelligent Parameter Selection.* Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, 2005: p. 4838-4841.

92. Sima, C., U. Braga-Neto, and E.R. Dougherty, *Superior feature-set ranking for small samples using bolstered error estimation.* Bioinformatics, 2005. **21**(7): p. 1046-1054.

93. Huang, E., et al., *Gene expression predictors of breast cancer outcomes.* The Lancet, 2003. **361**(9369): p. 1590-1596.

94. Wu, Z. and R.A. Irizarry, *Preprocessing of oligonucleotide array data.* Nature Biotechnology, 2004. **22**(6): p. 656-658.

95. Hess, K.R., et al., *Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer.* Journal of Clinical Oncology, 2006. **24**(26): p. 4236.

96. Amin, M.B., et al., *Prognostic Impact of Histologic Subtyping of Adult Renal Epithelial Neoplasms: An Experience of 405 Cases.* The American Journal of Surgical Pathology, 2002. **26**(3): p. 281.

97. Kaufmann, V. and R. Ladstädter, *Elimination of color fringes in digital photographs caused by lateral chromatic aberration.* Proceedings of the XX International Symposium CIPA, 2005. **26**: p. 403-408.

98. Phan, J.H., et al., *Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment.* Trends in Biotechnology, 2009. **27**(6): p. 350-358.

99. Paterlini-Brechot, P. and N.L. Benali, *Circulating tumor cells (CTC) detection: Clinical impact and future directions.* Cancer letters, 2007. **253**(2): p. 180-204.

100. Riethdorf, S., et al., *Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system.* Clinical Cancer Research, 2007. **13**(3): p. 920.

101. Ghossein, R.A., et al., *Detection of circulating tumor cells in patients with localized and metastatic prostatic carcinoma: clinical implications.* Journal of Clinical Oncology, 1995. **13**(5): p. 1195.

102. Ghossein, R.A. and J. Rosai, *Polymerase chain reaction in the detection of micrometastases and circulating tumor cells.* Cancer, 1996. **78**(1).

103. Wharton, R.Q., et al., *Increased Detection of Circulating Tumor Cells in the Blood of Colorectal Carcinoma Patients Using Two Reverse Transcription-PCR Assays and Multiple Blood Samples 1*. 1999, AACR. p. 4158-4163.

104. Molnar, B., et al., *Circulating tumor cell clusters in the peripheral blood of colorectal cancer patients.* Clinical Cancer Research, 2001. **7**(12): p. 4080-4085.

105. Liotta, L.A., M. Ferrari, and E. Petricoin, *Clinical proteomics: written in blood.* Nature, 2003. **425**(6961): p. 905.

106. Petricoin, E.F., et al., *The blood peptidome: a higher dimension of information content for cancer biomarker discovery.* Nature Reviews Cancer, 2006. **6**(12): p. 961-967.

107. Petricoin, E.F., et al., *Use of proteomic patterns in serum to identify ovarian cancer.* The Lancet, 2002. **359**(9306): p. 572-577.

108. Baggerly, K.A., J.S. Morris, and K.R. Coombes, *Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.* Bioinformatics, 2004. **20**(5): p. 777-785.

109. Baggerly, K.A., et al., *Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer*. 2005, Journal of the National Cancer Institute, Vol. 97, No. 4,© Oxford University Press 2005, all rights reserved. p. 307-309.

110. Ransohoff, D.F., *Lessons from Controversy: Ovarian Cancer Screening and Serum Proteomics*. 2005, Journal of the National Cancer Institute, Vol. 97, No. 4,© Oxford University Press 2005, all rights reserved. p. 315-319.

111. Han, M., et al., *Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules.* Nature Biotechnology, 2001. **19**(7): p. 631-635.

112. Peterson, J.J. and T.D. Krauss, *Photobrightening and photodarkening in PbS quantum dots.* Physical Chemistry Chemical Physics, 2006. **8**(33): p. 3851-3856.

113. Camp, R.L., G.G. Chung, and D.L. Rimm, *Automated subcellular localization and quantification of protein expression in tissue microarrays.* Nature Medicine, 2002. **8**(11): p. 1323-1328.

114. McCabe, A., et al., *Automated quantitative analysis (AQUA) of in situ protein expression, antibody concentration, and prognosis.* jnci, 2005. **97**(24): p. 1808-1815.

115. Qian, X., et al., *In vivo tumor targeting and spectroscopic detection with surface-enhanced Raman nanoparticle tags.* 2007.

116. Kim, J.H., et al., *Nanoparticle probes with surface enhanced Raman spectroscopic tags for cellular cancer targeting.* Anal. Chem, 2006. **78**(19): p. 6967-6973.

117. Vo-Dinh, T., L.R. Allain, and D.L. Stokes, *Cancer gene detection using surface-enhanced Raman scattering (SERS).* Journal of Raman Spectroscopy, 2002. **33**(7).

118. Siy, P.W., et al., *Matrix factorization techniques for analysis of imaging mass spectrometry data.* Proceedings of the 8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008, 2008: p. 1-6.

119. Cornett, D.S., et al., *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems.* Nature Methods, 2007. **4**: p. 828-833.

120. Schwartz, S.A., et al., *Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry.* Cancer Research, 2005. **65**(17): p. 7674-7681.

121. Chaurand, P., et al., *Imaging mass spectrometry: principles and potentials.* Toxicologic Pathology, 2005. **33**(1): p. 92-101.

122. Stoeckli, M., et al., *Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues.* Nature medicine, 2001. **7**(4): p. 493-496.

VITA

Richard A Moffitt

MOFFITT was born in Merritt Island, Florida. He received his primary education at Pine View School for the Gifted in Osprey, Florida, and received a B.S. in Biomedical Engineering from Georgia Institute of Technology in December of 2004. He then continued his studies at Georgia Tech in January of 2005 and began to pursue a doctorate in Bioengineering. When he is not working on his research, He enjoys spending time with family and friends while traveling to see his beloved Yellow Jackets play football.